Multimedia Tools and Applications (2023) 82:11265–11283
https://doi.org/10.1007/s11042-022-13463-1

1226: DEEP-PATTERNS EMOTION RECOGNITION IN THE WILD

A method for simplifying the spoken emotion recognition system using a shallow neural network and temporal feature stacking & pooling (TFSP)

Krishna Chauhan[1] · **Kamalesh Kumar Sharma**[1] · **Tarun Varma**[1]

Received: 1 November 2021 / Revised: 2 April 2022 / Accepted: 6 July 2022 /
Published online: 10 August 2022

Abstract

This study presents a new speech emotion recognition (SER) technique using temporal feature stacking and pooling (TFSP). First, Mel-frequency cepstral coefficients, Mel-spectrogram, and emotional silence factor (ESF) are extracted from segmented audio samples. The normalized features are fed into this neural network for training. For final feature representation, the learned features passed through the proposed TFSP framework. Subsequently, a linear support vector machine classifier is employed for emotion classification. It is evident from the confusion matrices that the suggested method can extract emotional content from speech signals efficiently with more unique emotional aspects from commonly confused emotions. According to this study, a shallow neural network can perform as good as the existing deep learning architectures like CNN, RNN, and attention networks. It may be mentioned here that the proposed method also utilises data augmentation by artificially increasing the number of speakers by disrupting the vocal tract length. Furthermore, these highly complex networks employ millions of trainable parameters, resulting in a longer convergence time.

The experiments are carried out on four different language speech emotional datasets, the Berlin emotional speech dataset (EmoDB) in German language, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) in North American English, Surrey Audio-Visual Expressed Emotion Database (SAVEE) in British English and a newly constructed MNITJ-Simulated Emotional Hindi speech Database (MNITJ-SEHSD) in the Hindi language. Experimental results on the proposed framework achieved an overall accuracy of 95.09%, 90.20%, 95.50% and 94.67%, on EmoDB, RAVDESS, SAVEE and MNITJ-SEHSD, respectively, at much lesser computational complexity. These findings are compared to the baseline of the three existing architectures on the same databases. Classification accuracy, precision, recall and F1-score are used to validate the developed method.

Keywords Speech emotion recognition · Vocal-tract length disruption · SVM classifier · Average pooling · Shallow neural network

✉ Krishna Chauhan
2018rec9121@mnit.ac.in

Extended author information available on the last page of the article.

1 Introduction

It is well known that detecting a person's emotional condition is a challenging task. In an identical situation, each person communicates their feelings uniquely. Automatically detecting these emotional states of human speech is difficult for machine learning systems. Though there are other modalities for detecting the emotional state of humans, recently, speech-based emotion recognition is gaining even more popularity, specifically in deep learning models [2, 17]. Due to the widespread use and development of human-computer interface systems, speech emotion identification has recently gained popularity as a research topic in signal processing, artificial intelligence, and other fields [13]. Our voice is the most natural, fastest, and most widely used mode of communication when it comes to modes of communication. Emotional recognition in speech improves the effectiveness of detection for applications such as speech to text, as well as strengthens the personification of machines [17]. Two primary directions for investigating speech emotions are dimensional and categorical. Categorical emotion recognition is more typically utilized under the assumption that a few basic emotions can be recognized universally, such as happiness, anger, sadness, surprise, fear, and disgust. In the dimensional category, the emotions are classified based on how positive, hostile, strong, weak, calm, or excited the emotions are [2, 40].

For researchers, the most important aspect is how to extract the powerful, discriminating and emotionally rich acoustic contents of the speech signals. The initial development of speech emotion recognition(SER) systems was based on hand-crafted features from the audio signal, for example, Mel Frequency cepstrum coefficients (MFCC), Linear predictive coefficients (LPC), Teager energy-related features (TEO), Fourier parameters (FP) etc. [4, 15, 43]. Though the results were reasonably good, the efficiency is always inferior to that of humans. These features are low-level audio representations and may not be discriminant enough to classify speech-based emotions. Various deep-learning algorithms have been used to tackle this problem by extracting high-level audio features. Recurrent neural networks (RNNs) have proven greater performance in recognizing voice emotion as they can learn the sequential data very well, such as speech. In RNNs, the problem of considering long-term dependencies has been mitigated by using long-short-term memory (LSTM), and temporal convolutional networks (TCNs), which was due to their limited capturing of temporal data [33, 39, 49]. TCNs can not copy the long-term dependencies between long-range audio patterns, and this issue further leads to a bottleneck of classification accuracies in these models. Other solutions are provided by using bi-directional analysis, which can further better track the past information of the sequences and eventually make the structure more robust in [32]. Convolutional Neural Networks (CNN) are the Deep Neural Networks (DNN) extension that addresses challenges such as determining the best features to use as inputs to these neural networks for effective learning. Aside from their tremendous success in visual tasks [25], some recent studies have successfully used CNNs for feature learning in speech signal processing. Feature learning in speech emotion recognition has been effectively applied using a variety of CNN architectures [8, 18]. Also, CNN and LSTM have been employed in many studies throughout the last decade. They are, nevertheless, constrained by the lack of sequence parallelization and high computation times. Meanwhile, the attention mechanism is used in learning significant feature representations for SER. Self-attention is used to capture the contextual dependencies specifically for NLP tasks, but more recently they have achieved state of art in SER [4, 11, 47].

The above mentioned different architectures and their combinations are used to enhance the classification accuracies of emotion identification. The use of diverse learning models to improve these accuracies often comes at the expense of making the overall architecture more complex. Nonetheless, the number of trainable parameters has increased due to the structures like attention mechanisms used in the models. If the dataset is small, then certainly it is not adequate. Hence even in the presence of significant computational resources, performance is only tolerable because of the extensive input features and complexities of modern deep learning models. This work proposes a modified neural network design for better temporal feature learning to address this issue.

2 Related work

An effective SER system is comprised of two essential components: relevant feature extraction and emotion categorization. This part covers a quick overview of emotion categorization using speech based on complex current architectures to emphasize the suggested design simplicity, which is more relevant to our work.

Spectrum characteristics are one of the paralinguistic features commonly employed in speech emotion recognition. These feature representations are classified using several classifiers in past such as SVM, GMM and HMM [36, 41]. Machine learning techniques boost the progress of emotion identification using speech signals compared to the hand-crafted audio features classification. In the case of CNNs, the features are better learned because they can capture detailed time-frequency correlations than a dense network. Han et al. used a DNN inputted with hand-crafted features from speech segments to create utterance level SER features, which are fed into an Extreme Learning Machine (ELM) for classification [19]. The statistical values of the entire sequence are classified using ELM, and they have observed a net improvement of 5% compared to the SVM classifier. In our previous work, a compact five-layer CNN network achieved 84.61% on EmoDB with log-Mel-spectrogram processed as an image [10]. This only five-layer network produces promising results on observed datasets. The performance of a standard SVM is compared to MLP classifiers using five different speech representations in [24]. On RAVDESS, they achieved a weighted accuracy of 78.8% using data augmentation with SMOTE (Synthetic Minority Over-sampling Technique). Anvarjon et al. suggested an SER model with minimal computing complexity and high recognition accuracy based on only CNN layers, which was then classified using SVM [3]. Trigeorgis et al. proposed merging convolutional neural networks (CNNs) with LSTM networks to automatically learn the best representation of the speech signal straight from the raw temporal representation to solve the problem of context-awareness [42]. Chen et al. proposed another CNN-LSTM architecture to improve state of the art, having 3-D Mel-spectrogram as input to the model [11]. Mel-scaled spectrograms and their two differentials, deltas and double deltas, are used in their experiments. Calculating the deltas and double-deltas for individualized attributes is considered to maintain useful emotional information while reducing the influence of emotionally irrelevant components, resulting in less misclassification. Zhang et al. used a pre-trained AlexNet model to construct an SER model to extract deep features from audio spectrograms using transfer learning. A fully connected neural network is integrated with an attention mechanism to forecast the embedded emotion [48]. Nediyanchath et al. used a multitask learning transformer-based network with gender identity for emotion identification. They have also incorporated positional embeddings of emotionally relevant cues for final classification. Their architecture can achieve the state of the art results on a challenging emotional dataset like IEMOCAP [35]. Peng et al. employed 3-dimensional convolution to incorporate a time-frequency scale with BiLSTM and attention weighing. This structure reinforces the time step importance by using attention weights. The series continues with more advanced structures such as autoencoders [16], capsule networks [46], generative adversarial networks [9], and various convolutional neural network modifications. In this paper, a shallow neural network is explored to identify emotions in human speech by processing acoustic features. The summary of the primary contributions of the proposed work is as follows:

(i) The audio samples are segmented to increase the number of samples available with respective source labels for training the model.
(ii) For the network input, only two spectral features, MFCC and Mel-spectrogram, are utilized, and one temporal feature, the emotional silence factor, is proposed, which proved effective in the final classification.
(iii) To conduct utterance level analysis from segments, a multi-layer feature learning is proposed using the TFSP.
(iv) The end-to-end architecture performs reasonably well with the most state-of-art findings and helps in decreasing both the cost of computation and the time complexity.

The relevant research and some current findings in the field are discussed in Section 2. The remaining portion of the paper is organised as follows: The proposed methodology with the database summary is discussed in Section 3. The performance evaluation, results, and discussion are addressed in Section 4, followed by a conclusion in Section 5.

3 Methodology

3.1 Database

This study employed four emotional speech databases; namely, Malaviya National Institute of Technology Jaipur Simulated Emotional Hindi Speech Database (MNITJ-SEHSD), the Berlin Emotional Speech Database (EmoDB) [6], Surrey Audio-Visual Expressed Emotion Database (SAVEE) [22], and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [29].

3.1.1 MNITJ-SEHSD database

A new database is recorded for this work, MNITJ-SEHSD in the Hindi language at Malaviya National Institute of Technology Jaipur. The speech corpus is designed by imitating five different emotions with neutral (non-emotional) text, allowing the speakers to perform and mimic the feelings without bias. The emotions present in the database are neutral, angry, happy, fear and sad. The audio samples were collected from ten speakers, five males and five females. The speakers are between the ages of 19 and 27. An omnidirectional microphone is used to record these emotive audio recordings, and its position can be changed or adjusted as needed. The Omni microphone is always a good choice when sound must be detected and captured from multiple angles or regions. Background noise and reverberation did not affect the recording as done in a quiet environment. The microphone input volume is set accordingly, and data recordings are made at a sampling rate of 44.1 kHz. Each speaker uttered ten emotionally neutral sentences in each of the five emotions. As a result, each speaker made 50 utterances, resulting in 50 X 10 = 500 utterances in the dataset. The recording period of an emotional utterance varies with speakers and ranges from 2.5 to 4 seconds. The number of words in each sentence is six except three sentences, two with five words and one with seven words. The distance between the microphone and the speaker is about 50 centimetres.

All the recordings are done speaker-wise, emotion by emotion. Three experts evaluated the authenticity of the corpus. Ten predetermined Hindi sentences are chosen for recording, as follows:

1. मोहन अपनी किताब पढ़ता है (Mohan reads his book.)
2. तुमने कल रात में मैच खेला (You played the match last night.)
3. बच्चे आज तुम्हारे घर जाएंगे (Kids will go to your house today.)
4. मेज़ पर कपडा पड़ा हुआ है (The cloth is lying on the table.)
5. राम ने सीता को पत्र लिखा (Ram wrote letter to Sita.)
6. कल हमारे घर में शादी है (There will be wedding tomorrow in our house.)
7. हम आज क्रिकेट खेल रहे है (We are playing cricket today.)
8. मैं अपना काम कर रहा हूँ (I am doing my work.)
9. हम अपना नाश्ता कर रहे है (We are having our breakfast.)
10. आज रात को बारिश होने वाली है (It’s going to rain tonight.)

3.1.2 EmoDB database

EmoDB includes speech samples from five professional actors and actresses, pronounced in ten distinct sentences and representing seven emotions as anger, boredom, disgust, fear, happiness, sadness and neutral [6]. The length of 535 audio recordings in German varies from 2 to 8 seconds. There were various situations under which the recording was done, like TV programs, interactive game shows, and talk shows. As per the gender information, the dataset is balanced, but it is an imbalanced dataset in the emotional category. Data augmentation is performed to artificially populate the data samples to address this problem.

3.1.3 RAVDESS database

It has a total of 24 actors, with equal male and female numbers [29]. There are eight emotions in the dataset: neutral, calm, happy, sad, angry, disgust, surprised and fearful uttered in North American English. It has 60 trials for each of the 24 actors, thus 1440 audio files. The dataset includes audio and video versions; however, because the goal is to distinguish emotions from speech alone, only audio data is used in experiments to compare our findings with the past models. RAVDESS is a relatively young dataset (created in 2018), with only two sentences per emotion and few words utilized in these two statements. All emotions have two levels of intensity: normal and intense. As the sentences are not extremely varied, this dataset’s classification accuracy is low compared to others in our network and past research. All 1440 wave files were acquired at a sampling rate of 48 kHz and encoded in 16 bits. A number sequence identifies the emotional label and speakers in these audio filenames.

3.1.4 SAVEE database

This emotional speech corpus is also an acted emotional speech dataset and has only four male actors who uttered seven emotions, i.e., anger, disgust, fear, happiness, neutral, sadness and surprise [17]. It has 480 audio files in the British English language. The sentences are taken from the TIMIT corpus. Although the dataset contains visual content, only audio files are used in the experiments. The sample rate is 44.1 kHz, and the encoding system is 16 bits, similar to RAVDESS. Ten experts evaluated the emotional corpus to ensure its quality.

Multiple datasets in various languages and recording situations test the proposed model. Table 1 summarises the key features of the four datasets. All of which provide emotional recordings of anger, happy, neutral, and sad emotions in common.

3.2 Audio augmentation

Vocal tract length perturbation and other audio augmentation techniques are frequently employed for speech-related speaker-independent applications such as speech recognition, and music genre categorization [1, 34]. One of the apparent hurdles in deep learning model training is less labelled data availability. However, a small neural network is not as data-hungry as the deep learning models but needs a fair amount of data samples to train the model. As more data samples are used to train the model, the recognition accuracy improves [21, 39], so to improve the classifier's generalization capability, audio samples are populated, which also deals with the problem of data sparsity in all three datasets. In experiments during training, vocal tract length perturbation is used as a means for increasing the number of speakers [23]. For every audio file, additional three replications are generated using nlpaug library [30]. The samples structure the algorithm's knowledge learned in the training set and test set to evaluate the model's performance. The final findings regarding classification accuracy indicate that the technique employed to augment the training samples artificially also does an excellent job of preserving the classes.

3.3 Pre-processing and generation of network input

The block diagram of the end to end process is given in Fig. 1. Pre-processing input audio samples is critical in establishing a model's performance for any signal processing activity. For model training, each utterance is divided into one-second chunks and ignored the ones that are too short. The segments are fed to the model as an individual utterances. The purpose is to gather more training data in the training set. Each of these audio segments is overlapped 60% with the previous one and labelled with its corresponding source label. In this way, the effective length is one second with 0.6 seconds overlapping. The spectral features are extracted for these chunks. Two spectral features are used: Mel frequency cepstral coefficients (MFCC) and the other is Mel- spectrogram. The third feature is the emotional silence factor (ESF), which measures the number of pauses or silences in a given utterance. The feature extraction process is detailed in the section below.

Table 1 Dataset summary used for the Experiments

Dataset	Number of speakers	Total utterances	Emotions
MNITJ-SEHSD	10 (5M, 5F)	500	Happy, Angry, Neutral,Sad, Fear
EmoDB [6]	10 (5M, 5F)	535	Angry, Boredom, Disgust, Fear, Happy, Sadness,Neutral
SAVEE [22]	4 (4M)	480	Angry, Disgust, Fear, Happy, Neutral, Sad, Surprised
RAVDESS [29]	24 (12M, 12F)	1440	Neutral, Calm, Happy, Sad, Angry, Disgust, Surprised

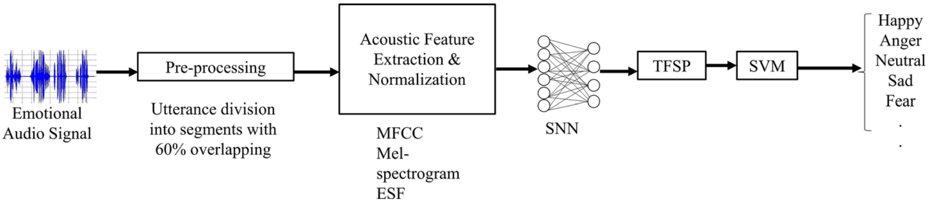
Emotional Audio Signal
Pre-processing
Utterance division into segments with 60% overlapping
Acoustic Feature Extraction & Normalization
MFCC
Mel-spectrogram
ESF
SNN
TFSP
SVM
Happy
Anger
Neutral
Sad
Fear
.
.

Fig. 1 An overview of the proposed speech emotion recognition system using shallow neural network (SNN) and TFSP

3.3.1 Mel frequency cepstral coefficient

In speech processing applications as sound classification and speech emotion identification, MFCCs are the most commonly used features [21, 50]. As humans do not perceive sound frequency linearly, Mel scaled characteristics compensate for this when dealing with speech signals. These characteristics mimic a human's natural sound frequency reception pattern. MFCC carries the information based on the physical properties of the speech signal. The following are the fundamental steps in extracting MFCC features: i) extracting windowed speech frames from raw audio signal ii) generating its spectrum using Discrete Fourier Transform iii) calculating its power spectrum iv) the linear frequency of the power spectrum is converted to the Mel frequency scale v) these frequencies are divided into uniform bands, and again converted band frequencies to their linear counterparts. vi) each of the bands is multiplied by the triangle window function vii) summing the energy of each of these triangular filters viii) calculating the logarithm of band energies and its discrete cosine transform leads to MFCC. A total of 40 MFCC coefficients were used in the experiments.

3.3.2 Mel spectrogram

It is another significant sound feature for identifying distinct audios. A two-dimensional representation of voice utterances is called a spectrogram. A spectrogram is also a superior choice that has been included in the majority of CNN deep models in the past and is still used to effectively classify human emotions [11, 21, 26].

3.3.3 Emotional silence factor

The speaker's emotion can influence the duration of pauses in human-to-human interaction; this feature measure silences or pauses in a given speech. For example, when a person is angry or joyful, they are less likely to pause, yet they may take longer pauses when sad or bored. The ESF is calculated as follows: (i) The hamming window is multiplied by each frame, then the energy is calculated. (ii) A frame's energy is given by:

$$E_f = \sum_{n=1}^{N} s_f^2(n) \tag{1}$$

where f is the frame number and N is the total number of samples in the f^{th} frame. (iii) the average energy of the corresponding segment will be:

$$E_{avg} = \frac{1}{M} \sum_{f=1}^{M} E_f^2 \tag{2}$$

where M is the total number of frames in that segment. (iv) An energy threshold is taken as $0.06 * E_{avg}$. Eventually, the frames having energy less than this threshold are considered silent frames. (v) Finally, ESF is defined as the ratio of silent frames to the total number of frames.

Figure 2 represents the waveforms of all seven emotional categories in the EmoDB database. In the second column, notice MFCC coefficients for angry and sad emotions; only low frequencies are seen in sad emotions, while other high-frequency amplitude values are not visible. However, more high-frequency components are apparent in the plot for the angry emotion. Mel-spectrogram is a helpful visualization feature that uses the Mel scale to visualize sound. Spectrogram allows us to focus on the most critical aspects of the audio. In the last column of Fig. 2, Mel-spectrograms of all emotions are shown for the same audio length as for MFCC.

3.4 Details of the architecture

It is a feedforward artificial neural network that converts input features into a set of valuable outputs. There are three layers: an input layer that receives the input signal, an output layer that makes the final judgments, and a hidden layer between the input and output layers. If the inputs are specified as x, the weights as w and the bias as b, then the network's output is defined as:

$$z = ReLU(\mathbf{wx} + \mathbf{b}) \tag{3}$$

where z is the final output or the input of another layer. An activation function is also required for introducing nonlinearity to the network, as in real-world applications. ReLU activation is utilized for the same. It is defined as:

$$ReLU(a) = max(0, a) \tag{4}$$

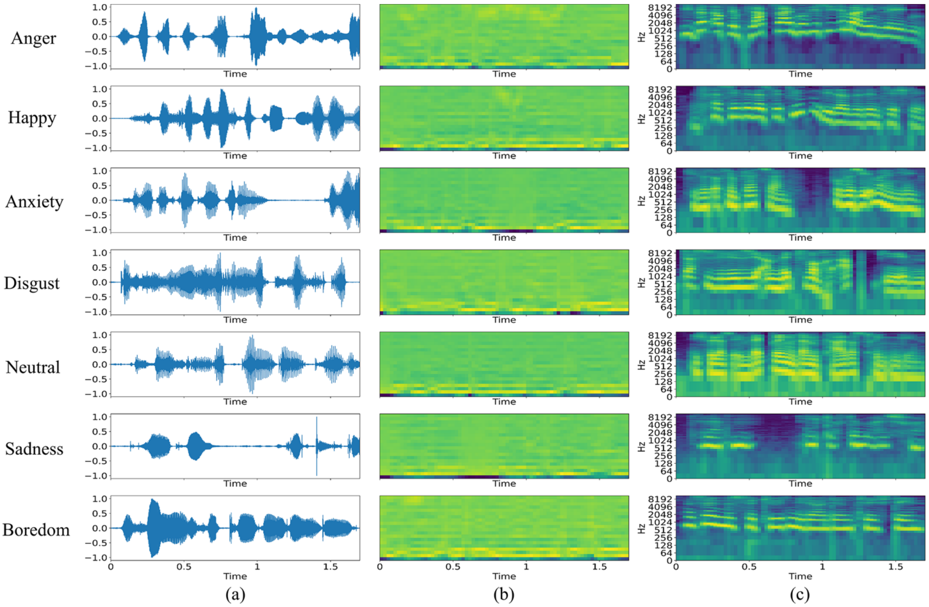

Fig. 2 The audio waveforms of seven emotions from EmoDB database (a) The audio signal (b) MFCC (c) Mel-spectrogram

where $a = \mathbf{wx} + \mathbf{b}$. Sparseness and a lower probability of vanishing gradient are two important advantages of ReLUs, as the gradient is less likely to collapse. Vanishing gradient occurs for more layers and when $a > 0$. Due to the continual gradient of ReLUs, learning is accelerated. The output of the fully connected layer is computed as:

$$y_k = f\left(\sum_l \mathbf{W}_{kl}\mathbf{x}_l + \mathbf{b}_k\right) \tag{5}$$

where f is the activation function (ReLU in this case), y_k denotes the k^{th} output neuron, x_l is the l^{th} input neuron, W_{kl} is the weight parameter transforming input to output and b_k is the corresponding bias term.

The input for the neural network is a feature vector of dimensions $I \in R^{SXd}$, where S represents the number of audio segments, and d is the dimension of the audio features. The network makes use of 40-dimensional MFCC vectors and 128 bands of Mel-spectrogram. ESF is a vector with only one dimension. In the hidden layer, 500 neurons are used for network processing. In speech emotion recognition tasks, raw audio, when processed directly by complicated neural network architectures like CNN-LSTM-Attention, produces higher results [39]. In contrast, simpler architectures such as feedforward neural networks require a better representation of audio data as input for better classification. The network classifier trains continuously as, after every step, the partial derivative of the loss function is calculated for the current model parameters to evaluate the updated model parameters.

According to the hidden layer's calculation, the output layer classifies the input features and assigns the expected emotion label. The model is developed using the Pytorch framework with librosa to extract audio features. Because of its higher convergence rate, the Adam optimizer is used. The L2 regularization factor of 0.01 is chosen with an adaptive learning rate for all experiments. This regularization term is added to the loss function, which reduces the model parameters to prevent overfitting. The adaptive learning rate maintains the initial rate of 0.001 until the loss continues to decrease.

Finally, the model with the best outcome is adopted. The batch size is 128 for all four datasets, and the Adam optimizer's numerical stability factor is set to 1e-08. Additionally, a feature set from the openSMILE toolkit, the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), is used with the proposed model. The results are not as good as those obtained using the acoustics feature set, as shown in Table 7.

3.5 TFSP for effective feature representation

On completion of the model training, the output feature of the hidden layer is stored as $X = (x_1, x_2, x_3, ..., x_N) \in R^{FXS}$ with F as the feature dimension and S as the number of overlapping segments in the utterance. As the number of segments may vary in an audio sample, these segment features are not helpful for speech emotion identification tasks. The segment features need to be converted into utterance features so that the final dimension is the same for all the utterances. Pooling strategies are used in the past for the same, like mean and max pooling. The pooling of these segmented features must be chosen such that the maximum temporal information is intact for final classification.

Before pooling operation, the extracted segment features are divided into some fixed number of blocks, as shown in Fig. 3. A similar strategy is used in [27] for partitioning of the image into increasingly fine sub-regions. These blocks are non-overlapping and segmented along the time axis. Segmentation up to level four is used for the proposed model as the average length of the audio samples is 4 seconds for three datasets. For RAVDESS, segmentation of up to five levels is used. For example, only one block is used for level one, so there

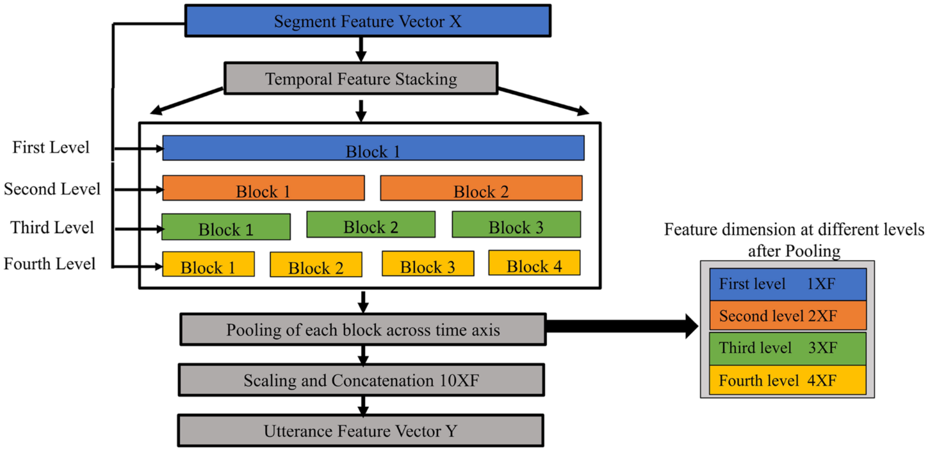

Fig. 3 The frameowrk of Temporal Feature Stacking & Pooling (TFSP)

is no segmentation; for level two, the segment feature vector is divided into two equal non-overlapping blocks and so on. Each block will have a dimension of FXs with s segments in that block; average pooling is utilized to produce stationary length features for all the utterances. The features, after pooling, encode all temporal information at different levels. The higher the level more polished temporal information is restored. For example, compared to level 1 (which has the average of all segments features along the time axis), level 4 embeds more segmented block's average values and hence more precise temporal cues.

Accordingly, the features at all levels are scaled; the explanation is as follows: At level one, the pooled features are scaled by 4 (as there is a total of 4 levels), at level two by 3, at level three by two and level four by 1. As level 4 consists of maximum temporal information, it should be given the highest weightage. These scaled features are now concatenated at all levels to get the SVM classifier's input feature for a given utterance. The pooled features at different levels are transposed as shown in Fig. 3 to make them compatible with SVM input.

The essence of the whole TFSP process is; the segmented audio feature vector X is divided into several blocks at different levels (for example, 4). Each block's features are passed through the pooling mechanism, and after an optimum scaling, features from all levels are concatenated to convert the segment feature vector into an utterance feature vector, such that the dimension for each utterance is identical for further processing in this case, $10XF$. Similar to TFSP, scaling techniques have been used in [27]. In the following section, the verification of the validity of the proposed scheme is presented through experimental results on all four emotional speech datasets.

3.6 Multiclass support vector machines

Support vector machines (SVM) are supervised machine learning approaches that outperform other classifiers in the majority of pattern recognition experiments [12]. The SVM classifier has been utilised in the majority of emotion recognition studies [5, 7, 44]. There are two techniques for solving a multi-label classification problem with SVM known as the "one-versus-on" strategy and "one-versus-all". Because it is more reliable in practice, the one-versus-one strategy was adopted here for the experiments [20]. The input features to be classified are transformed into a high-dimensional space using any kernels available, such as linear & nonlinear functions, radial basis functions, polynomial functions, and sigmoid functions. An optimum hyperplane is chosen in the modified feature space for division into the classes. As the SVM is a binary classifier, a mixture of binary SVM classifiers has to be used in multiclass problems, and linear kernel SVM classifiers are employed in a one-against-one approach. The database is partitioned into five mutually exclusive subsets for training and validation, according to the 5-fold cross-validation scheme. The classifier is trained and tested five times, with one set serving as the testing set and the other four as the training set every time.

4 Performance evaluation, experimental result, and discussion

The classification accuracies and F1 scores for each emotion category are presented in this section. The experiments are designed for speaker-dependent analysis, i.e., the training and test utterances are not dependent on speakers. The performance of the proposed technique is evaluated using two statistical parameters: weighted and unweighted accuracy. Weighted accuracy is a measure of classification performance throughout the test set, and unweighted accuracy is an average of recall for each emotional class, which is a more appropriate metric for the experiments given the highly uneven nature of emotions distribution in the EmoDB datasets. Training the model involves randomness due to the distribution of data samples for training and testing. For the experiments, random seeds are utilized to strengthen the robustness of the results. To better analyze the statistical parameters on each dataset, Recall, Precision and F1-score are also evaluated to show the classification summary for every class. Apart from the categorical classification of emotional classes, analysis of the arousal dimension of emotions is also performed on all four datasets. This dimension signifies the strength of the mood. For example, sad is a negative arousal emotion, whereas angry is a positive arousal emotion. The proposed SER framework can distinguish between emotions with high arousal, such as happiness, disgust, and anger, and emotions with low arousal, such as neutral, calm, and sadness, with good accuracy. Figures 4 to 7 depicts the classification results using confusion matrices of all emotions present in datasets. The true predicted accuracies are shown in the main diagonal of the matrix.

Figure 4 shows classification results on the newly constructed Hindi dataset, MNITJ-SEHSD, where ‘Neutral’, ‘Angry’, ‘Happy’ and ‘Fear’ can be recognized with accuracies of 97%, 96%, 95% and 96% respectively. The emotional class ‘Sad’ is identified with an accuracy of 90%. The average classification accuracy is 94.67%, and the unweighted classification accuracy is 94.8%. The average classification with high arousal classes as ‘Angry’, ‘Happy’ and ‘Fear’ is 95.5% and for ‘Neutral’ and ‘Sad’ is 93.4%. The identical classification of all classes is due to an equal amount of utterances in each category. However, the model cannot recognize the ‘sad’ class with greater than 90% accuracy, while all other emotions are identified with more than 95% on this dataset.

In Fig. 5 it is illustrated that on the EmoDB dataset, ‘Sadness’ is classified with the highest accuracy of 100%, and the other six emotions are classified with accuracies higher than 92%. The weighted accuracy for this dataset is 95.09% and the unweighted accuracy of 95.3%. The misclassifications between high and low arousal emotions are also very low; for example, anger is only mismatched with fear and happiness, while happiness is mismatched with anger. The high (Anger, Disgust, Fear and Happiness) and low (Boredom, Sadness and Neutral) arousal recognition accuracies are 95.10% and 94.56% on this dataset.

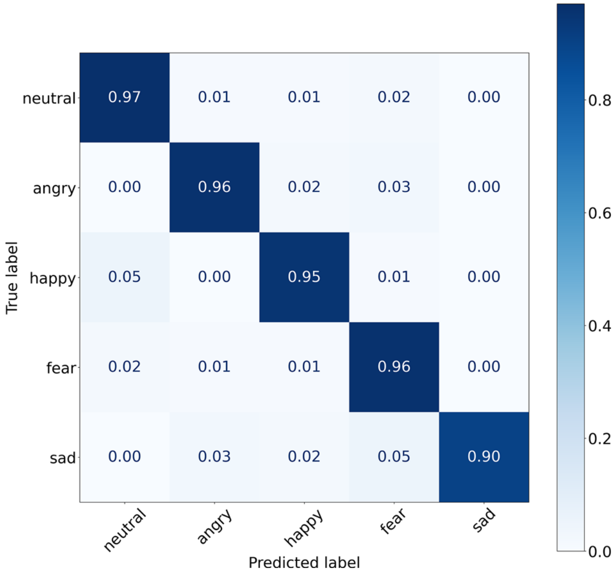

Fig. 4 Confusion Matrix of MNITJ-SEHSD Database

‘Neutral’, ‘Calm’, ‘Happy’ and ‘Angry’ emotions are distinguished with accuracy higher than 92% on the RAVDESS dataset as shown in the confusion matrix in Fig. 6. ‘Sadness’ is identified with the least accuracy of 86% similar to MNITJ-SEHSD and maximally misclassified with class ‘Neutral’, ‘Disgust’ and ‘Surprised’ in order. Here the lowest accuracy in this class is due to the least number of samples available (half compared to other emotional

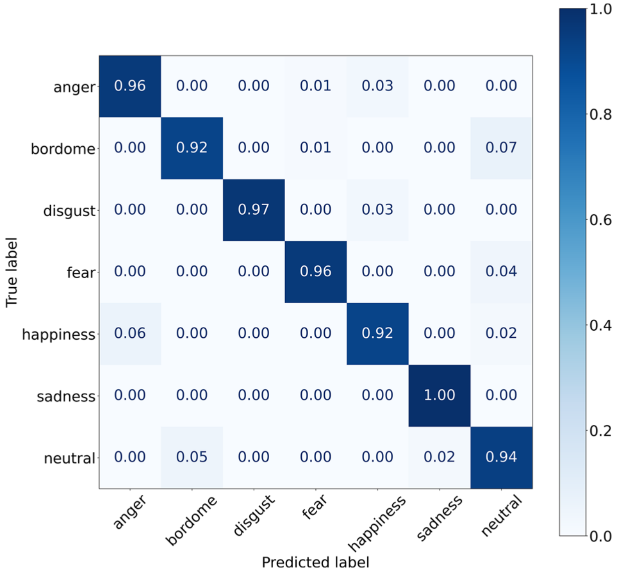

Fig. 5 Confusion Matrix of EmoDB Database

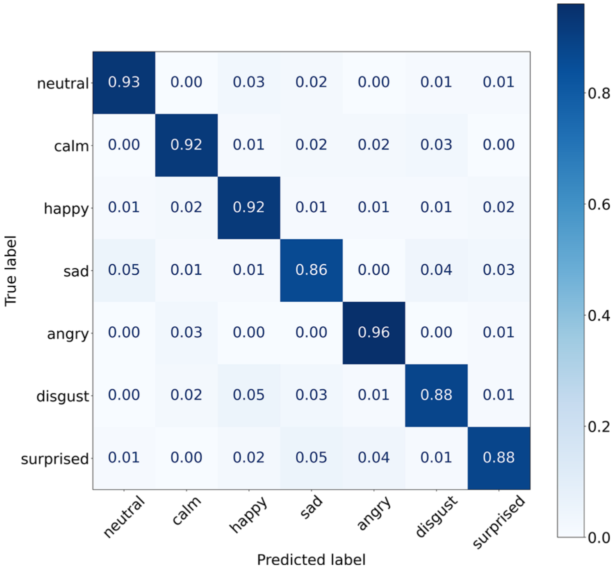

Fig. 6 Confusion Matrix of RAVDESS Database

classes). The proposed design observes the least performance on RAVDESS compared to the other datasets in the experiments; this could be due to the different language and culture. The average classification for the whole dataset is 90.2%, and the average recall for each class is 90.7%. The classification accuracies on high and low arousal dimensions are

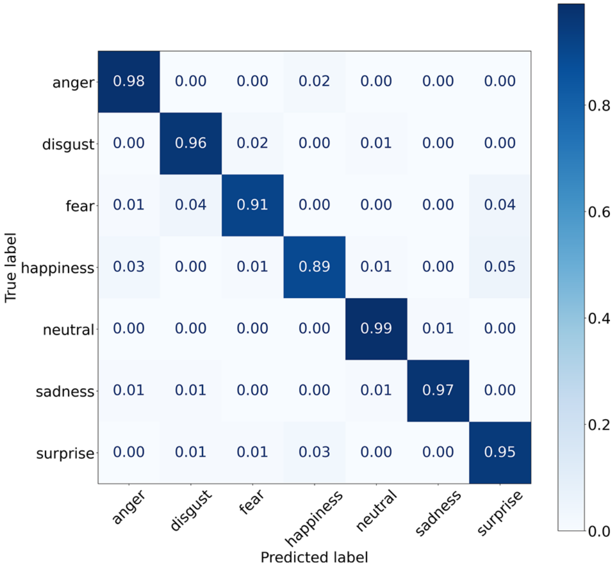

Fig. 7 Confusion Matrix of SAVEE Database

Table 2 The proposed method performance using MNIT-SEHSD Dataset

Emotion	Precision	Recall	F1 Score
Neutral	0.94	0.97	0.95
Angry	0.95	0.96	0.95
Happy	0.95	0.95	0.95
Fear	0.90	0.96	0.93
Sad	1.00	0.90	0.95

Table 3 The proposed method performance using EmoDB Dataset

Emotion s	Precision	Recall	F1 Score
Anger	0.97	0.96	0.97
Boredom	0.96	0.92	0.94
Disgust	1.00	0.97	0.99
Fear	0.96	0.96	0.96
Happiness	0.92	0.92	0.92
Sadness	0.98	1.00	0.99
Neutral	0.88	0.94	0.91

Table 4 The proposed method performance using RAVDESS Dataset

Emotion	Precision	Recall	F1 Score
Neutral	0.94	0.93	0.93
Calm	0.94	0.92	0.93
Happy	0.90	0.92	0.91
Sad	0.87	0.86	0.86
Angry	0.85	0.96	0.90
Disgust	0.89	0.88	0.88
Surprised	0.92	0.88	0.90

Table 5 The proposed method performance using SAVEE Dataset

Emotion	Precision	Recall	F1 Score
Anger	0.94	0.98	0.96
Disgust	0.94	0.96	0.95
Fear	0.95	0.91	0.93
Happiness	0.96	0.89	0.92
Neutral	0.98	0.99	0.99
Sadness	0.99	0.97	0.98
Surprise	0.92	0.95	0.93

Table 6 Observed Accuracy scores for all the datasets under examination

Dataset	MNITJ-SEHSD	EmoDB	RAVDESS	SAVEE
Accuracy(%)	94.67	95.09	90.20	95.50

90.2% and 90.1%, respectively. However, the model shows better results with some complex architectures, as shown in Table 7.

Figure 7 indicates that only one emotion, i.e., 'Happiness', is recognized with accuracy less than 90% in the SAVEE dataset, while the remaining six emotions with accuracies higher than 91%. The emotional classes are also misclassified with similar classes as Anger with Happiness, Happiness with Anger and Surprise. This further leads to the higher dimensional accuracy of 93.3% for high arousal and 98.5% for low arousal emotional classes.

The classification reports in terms of precision, recall and F1 score are also given from Tables 2, 3, 4 and 5 for a better understanding of the performance of suggested work on all used speech datasets. Table 6 shows the final conclusive accuracy chart of all the used datasets.

4.1 Performance comparison with the state-of-art

The comparisons between the proposed model and previous works with more complicated architectures are shown in this part. All techniques are speaker-dependent; thus, the comparison is based on that assumption.

Table 7 Compariosn of classification accuracy with complex architectures (%) (MME=MFCC + Melspectrogram + Emotional silence factor)

Dataset	Method	Research Work	WA	Year
MNIJ-SEHSD	proposed	MME	**94.67**	2022
		GeMAPs	92.67	2022
EmoDB	CNN	Dias Issa [21]	86.10	2020
	CNN+LSTM	Jianfeng Zhao [49]	76.40	2019
	CNN+LSTM+Attention	Hao Meng [32]	90.78	2019
	Proposed	MME	**95.09**	2022
		GeMAPs	93.45	2022
RAVDESS	CNN	Dias Issa [21]	71.60	2020
	CNN+LSTM	Mustaqeem [38]	77.02	2020
	CNN+LSTM+Attention	A.Orhan [4]	96.18	2021
	Proposed	MME	**90.20**	2022
		GeMAPs	88.30	2022
SAVEE	CNN	ABA Qayyum [37]	83.61	2019
	CNN+LSTM	R.Dangol [14]	78.18	2020
	CNN+LSTM+Attention	A.Orhan [4]	87.50	2021
	Proposed	MME	**95.50**	2022
		GeMAPs	94.12	2022

Table 7 illustrates the comparison for three datasets used in experiments. The comparisons were made using three baseline network architectures: CNNs, their combination with LSTM, and recent advancements in these networks with attention modules. In a baseline CNN-LSTM architecture, the number of trainable parameters is 16,736,324 for the EmoDB dataset; in contrast, for the proposed method, it is only 88507 for training the primary neural network [49]. Further addition of any deep learning component to this baseline structure will increase the number of training parameters of the model, and hence, in comparison, the proposed model gives far fewer trainable parameters for classification.

The model's performance on EmoDB is showing a definite 5% improvement compared to the [32], which is the most complex deep architecture in Table 7. For the RAVDESS dataset, the results are better using the method presented in [4]. However, data augmentation techniques have not been performed. For comparison of the architecture, this work is included. Otherwise, the classification accuracy is much improved compared to the five-layer CNN architecture given by [21]. Finally in SAVEE emotional dataset, an absolute improvement of 8% over [4] with self-attention model.

5 Conclusion

This paper proposes a shallow neural network model with temporal feature stacking and pooling (TFSP) applied over the learned features of the trained network and classifies them using a support vector machine (SVM) for spoken emotion recognition. With the suggested TFSP mechanism, a basic shallow neural network can outperform several complex deep models such as CNN+LSTM and CNN+LSTM+ATTN. The network uses three different audio features: MFCC, Melspectrogram and the ESF, as input for speech emotion recognition. Compared to some of the above complex architectures with superior classification results, the proposed network has fewer learning parameters, indicating that the model is less complex and requires less training time. In a CNN+LSTM network, the number of learning parameters is enormous compared to the proposed network. Experimental results on three popular benchmark datasets, EmoDB, RAVDESS and SAVEE, demonstrate the effectiveness of this framework. Additionally, a newly designed Hindi language dataset is also utilized in the proposed method. Furthermore, the proposed method has been proven effective in capturing high versus low arousal classes.

In the future, besides acoustic features, the authors indent to improve the classification accuracy further using the linguistic and visual information [31, 45]. All the trials are for speaker-dependent frameworks, where the emotional state classification is independent of the speakers. Additional strategies for working with speaker-independent architectures, such as using more extensive data augmentation techniques, can also be practised [28]. A cross-language analysis is also in the future scope of this work, as four different languages are used to evaluate the model's performance.

Acknowledgements The authors would like to acknowledge the support of the ADSIP Laboratory at MNIT, Jaipur, for providing all the facilities for carrying out the work.

Author Contributions Krishna Chauhan: Conceptualization, Methodology, Python implementation, Investigation, Validation, Writing - original draft, Writing - review & editing. Kamalesh Kumar Sharma: Supervision, Project administration, Investigation, Validation, Review & editing. Tarun Varma: Supervision, Project administration, Investigation, Validation, Review & editing.

Funding The authors received no financial support for their research, writing, or publication of this paper.

Declarations In research involving human participants, all procedures were carried out in compliance with ethical guidelines.

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Aguiar RL, Costa YM, Silla CN (2018) Exploring data augmentation to improve music genre classification with convnets. In: 2018 international joint conference on neural networks (IJCNN). IEEE, pp 1–8
2. Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Comm 116:56–76
3. Anvarjon T, Kwon S et al (2020) Deep-net: a lightweight cnn-based speech emotion recognition system using deep frequency features. Sensors 20(18):5212
4. Atila O, Şengür A (2021) Attention guided 3d cnn-lstm model for accurate speech based emotion recognition. Appl Acoust 182:108260
5. Atmaja BT, Akagi M (2021) Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm. Speech Comm 126:9–21
6. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of german emotional speech. In: Ninth European conference on speech communication and technology
7. Calvo RA, D'Mello S (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput 1(1):18–37
8. Chatterjee R, Mazumdar S, Sherratt RS, Halder R, Maitra T, Giri D (2021) Real-time speech emotion analysis for smart home assistants. IEEE Trans Consum Electron 67(1):68–76
9. Chatziagapi A, Paraskevopoulos G, Sgouropoulos D, Pantazopoulos G, Nikandrou M, Giannakopoulos T, Katsamanis A, Potamianos A, Narayanan S (2019). In: Interspeech, pp 171–175
10. Chauhan K, Sharma KK, Varma T (2021) Speech emotion recognition using convolution neural networks. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). IEEE, pp 1176–1181
11. Chen M, He X, Yang J, Zhang H (2018) 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process Lett 25(10):1440–1444
12. Chen L, Mao X, Xue Y, Cheng LL (2012) Speech emotion recognition: Features and classification models. Digit Signal Process 22(6):1154–1160
13. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. IEEE Signal Proc Mag 18(1):32–80
14. Dangol R, Alsadoon A, Prasad P, Seher I, Alsadoon OH (2020) Speech emotion recognition usingconvolutional neural network and long-short termmemory. Multimed Tools Appl 79(43):32917–32934
15. Deb S, Dandapat S (2018) Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification. IEEE Trans Cybern 49(3):802–815
16. Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: 2013 humaine association conference on affective computing and intelligent interaction. IEEE, pp 511–516
17. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit 44(3):572–587
18. Guizzo E, Weyde T, Leveson JB (2020) Multi-time-scale convolution for emotion recognition from speech audio signals. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6489–6493
19. Han K, Yu D, Tashev I (2014) Speech emotion recognition using deep neural network and extreme learning machine. In: Fifteenth annual conference of the international speech communication association
20. Hsu C-W, Lin C-J (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425
21. Issa D, Demirci MF, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing Control 59:101894
22. Jackson P, Haq S (2014) Surrey Audio-Visual Expressed Emotion (savee) Database. University of Surrey, Guildford
23. Jaitly N, Hinton GE (2013) Vocal tract length perturbation (vtlp) improves speech recognition. In: Proc ICML Workshop on Deep Learning for Audio, Speech and Language, vol 117

24. Javaheri B (2021) Speech & song emotion recognition using multilayer perceptron and standard vector machine. arXiv:2105.09406
25. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
26. Kwon S et al (2020) A cnn-assisted enhanced audio signal processing for speech emotion recognition. Sensors 20(1):183
27. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). IEEE, vol 2, pp 2169–2178
28. Liu Z-T, Li K, Li D-Y, chen L-F, Tan G-Z (2015) Emotional feature selection of speaker-independent speech based on correlation analysis and fisher. In: 2015 34th Chinese control conference (CCC). IEEE, pp 3780–3784
29. 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
30. Ma E (2019) nlpaug: data augmentation for NLP https://github.com/makcedward/nlpaug. Accessed 01 Nov 2021
31. Mansoorizadeh M, Charkari NM (2010) Multimodal information fusion application to human emotion recognition from face and speech. Multimed Tools Appl 49(2):277–297
32. 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
33. Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE:2227–2231
34. Mushtaq Z, Su S-F (2020) nvironmental sound classification using a regularized deep convolutional neural network with data augmentation. Appl Acoust 167:107389
35. Nediyanchath A, Paramasivam P, Yenigalla P (2020). In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7179–7183
36. Özseven T (2018) Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition. Appl Acoust 142:70–77
37. Qayyum ABA, Arefeen A, Shahnaz C (2019) Convolutional neural network (cnn) based speech-emotion recognition. In: 2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON). IEEE, pp 122–125
38. Sajjad M, Kwon S et al (2020) Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. IEEE Access 8:79861–79875
39. Sarma M, Ghahremani P, Povey D, Goel NK, Sarma KK, Dehak N (2018) Emotion identification from raw speech signals using dnns. In: Interspeech, pp 3097–3101
40. Schlosberg H (1954) Three dimensions of emotion. Psychol Rev 61(2):81
41. Sun L, Zou B, Fu S, Chen J, Wang F (2019) Speech emotion recognition based on dnn-decision tree svm model. Speech Comm 115:29–37
42. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, Zafeiriou S (2016) Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5200–5204
43. Wang K, An N, Li BN, Zhang Y, Li L (2015) Speech emotion recognition using fourier parameters. IEEE Trans Affect Comput 6(1):69–75
44. Wu S, Falk TH, Chan W-Y (2011) Automatic speech emotion recognition using modulation spectral features. Speech comm 53(5):768–785
45. Wu C, Huang C, Chen H (2018) Text-independent speech emotion recognition using frequency adaptive features. Multimed Tools Appl 77(18):24353–24363
46. Wu X, Liu S, Cao Y, Li X, Yu J, Dai D, Ma X, Hu S, Wu Z, Liu X et al (2019) Speech emotion recognition using capsule networks. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6695–6699
47. Xie Y, Liang R, Liang Z, Huang C, Zou C, Schuller B (2019) Speech emotion classification using attention-based lstm. IEEE/ACM Trans Audio, Speech, Language Process 27(11):1675–1685
48. Zhang Y, Du J, Wang Z, Zhang J, Tu Y (2018) Attention based fully convolutional network for speech emotion recognition. In: 2018 Asia-pacific signal and information processing association annual summit and conference (APSIPA ASC). IEEE, pp 1771–1775
49. 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

50. Zhong S, Yu B, Zhang H (2020) Exploration of an independent training framework for speech emotion recognition. IEEE Access 8:222533–222543

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Krishna Chauhan[1] · **Kamalesh Kumar Sharma**[1] · **Tarun Varma**[1]

Kamalesh Kumar Sharma
kksharma.ece@mnit.ac.in

Tarun Varma
tvarma.ece@mnit.ac.in

[1] Electronics and Communication Engineering Department, Malaviya National Institute of Technology Jaipur, Jaipur, Rajasthan 302017, India