Multimedia Tools and Applications (2022) 81:31107–31128
https://doi.org/10.1007/s11042-022-12886-0

Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network

Shalini Kapoor[1] · Tarun Kumar[2]

Received: 23 February 2021 / Revised: 24 January 2022 / Accepted: 10 March 2022 /
Published online: 8 April 2022

Abstract
Stress and anger are two negative emotions that affect individuals both mentally and physically; there is a need to tackle them as soon as possible. Automated systems are highly required to monitor mental states and to detect early signs of emotional health issues. In the present work convolutional neural network is proposed for anger and stress detection using handcrafted features and deep learned features from the spectrogram. The objective of using a combined feature set is gathering information from two different representations of speech signals to obtain more prominent features and to boost the accuracy of recognition. The proposed method of emotion assessment is more computationally efficient than similar approaches used for emotion assessment. The preliminary results obtained on experimental evaluation of the proposed approach on three datasets Toronto Emotional Speech Set (TESS), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Berlin Emotional Database (EMO-DB) indicate that categorical accuracy is boosted and cross-entropy loss is reduced to a considerable extent. The proposed convolutional neural network (CNN) obtains training (T) and validation (V) categorical accuracy of T = 93.7%, V = 95.6% for TESS, T = 97.5%, V = 95.6% for EMO-DB and T = 96.7%, V = 96.7% for RAVDESS dataset.

Keywords Speech emotion recognition · Convolutional neural networks · Deep learning · Emotion change detection · Spectrograms

✉ Shalini Kapoor
shalini_kapoor311@rediffmail.com

[1] Research Scholar, Dr. A.P.J Abdul Kalam Technical University, Lucknow, India

[2] Department of Computer Science & Engineering, Radha Govind Group of Institution, Meerut, India

1 Introduction

Stress is a psychological state in which the body's physical and mental balance gets disturbed due to physical, mental, or emotional strain [6]. Anger and stress are feelings that erupt due to displeasure hurting an individual's personal life, health, and interpersonal relation. Today stress is a global concern with more and more people experiencing stress in their daily lives. This necessitates a system that could continuously monitor and highlight our affective states such as anger and stress. Alert generated using such systems will enable us to take prompt and timely action to avoid the risks associated with these. Nowadays, there has been a rapid increase in the demand for smart and voice-driven interfaces due to the comfort and convenience offered by them. Voice is expected to be a preferred medium to interact with next-generation user interfaces, besides gesture, brain-computer, augmented reality, and other interfaces.

Low et al. [15] Speech is not just mere communication of messages but also carries a wealth of information related to mood, stress, psychological behavior, and mental health [12]. There are a large number of cues that specify that there are objectively measurable speech parameters that could reflect the emotional state of a person [7]. Emotions are expressed in speech using different speaking styles, tone of voice, rate of speech and intonation, etc. Speech is composed of two important components linguistic and paralinguistic components. A linguistic component is what is said while the paralinguistic component refers to how it is said. Speech not only contains the message but also contains information related to the speaker's affective state such as feelings and emotions. Emotions are expressed in speech using different speaking styles, tone of voice, rate of speech and intonation, etc. Even after extensive research carried out in the field of speech emotion recognition (SER) systems still researchers are facing lots of challenges and we are still quite far from desirable performance and accuracy. An emotional speech dataset that is ethnically diverse phonetically/prosodically balanced and big enough to cater to possible variations related to age, gender, culture, style, and language are required we are still lacking such database. Annotation of emotion in natural speech corpora is a time-consuming and difficult task so there is a need to develop a robust mechanism to annotate natural speech corpora. Till now, there is no consensus regarding optimal size of unit of analysis and best set of features. Most of the studies carried in this field are based on isolated speech while real time SER systems require continuous speech processing. Several times perception of emotions is entirely different from internal feelings so a mechanism to gaze at internal emotions by a deeper insight into the cognitive processes is required. Models that could differentiate between expressed and experienced emotions using the information gathered from some other source which provides deeper insight into the cognitive process needs to be developed. We still lack a compact feature set that could differentiate emotional patterns associated with particular emotions [2]. According to past research two approaches were used in past for SER traditional and deep learning. The traditional approach follows pipeline architecture with different steps such as feature extraction, feature selection, and classification. Each step needs to be optimized separately which is quite time-consuming and a cumbersome task. Optimal feature selection is a very crucial step in SER that has a high impact on the accuracy and performance of the SER system. The selection of optimal features requires expertise and domain knowledge.

Deep neural networks (DNN) classifiers have provided an elegant solution to the problem of optimal feature selection by bypassing it. The concept is to employ an end-to-end network that receives raw data as input and outputs a class label. It's not necessary to compute hand-

crafted features or to figure out which parameters are best for categorization. The network takes care of everything. Specifically, during the training phase, the network parameters (i.e., the weights and bias values provided to the network nodes) are tuned to behave as features efficiently splitting the input into the required categories. In comparison to conventional classification, this otherwise highly straightforward technique comes with substantially higher needs for labeled data samples. The adoption of deep learning (DL) techniques was a major turning point in SER. In a wide range of classification applications, supervised DL neural network models have been proven to outperform classical techniques, with the classification of images being particularly successful [22]. DL is becoming more important in the early detection of the new coronavirus (COVID-19). In many hospitals throughout the world, DL has become the primary technique for automatic COVID-19 categorization and detection utilizing chest X-ray pictures or other types of imaging [25]. It is used in the detection of skin disease, [4] plant diseases, [10] cervical cancer, [17] in disease classification, etc.

Due to the advent of CNN, various studies have shown the feasibility of using spectrograms for SER. The spectrogram is a 2D representation of speech signal carrying useful patterns related to the speaker's affective state and importantly, it maintains the signal in its entirety. Both traditional and deep learning approaches offer their advantages and disadvantages. Using the traditional approach for SER there is the possibility of biased feature selection or missing important features. Deep learning algorithm shows good performance on training dataset but sometimes show poor performance on other datasets. DNN requires the configuration of millions of parameters, each with intricate interrelationships, which are referred to as black boxes since the model's behavior is difficult to explain, even when the structure and weights are visible. A crucial impediment to fully exploiting the promise of DNNs for emotion identification is the unavailability of a sufficient number of emotion-labeled speech datasets, which makes training a deep network from scratch difficult. Traditional and deep learning-based techniques have clear trade-offs. Traditional algorithms are well-known, transparent, and optimized for performance and power efficiency, whereas DL provides higher accuracy and variety at the expense of a significant amount of computer resources.

Hybrid approaches traditional and deep learning techniques to get the best of both worlds. In the proposed work we have used a hybrid feature set obtained fusing handcrafted and deep learned features for emotion assessment. In the proposed work we have tried to assess three affective states anger, stress, and neutral. The datasets used in the proposed study are TESS, EMO-DB, and RAVDESS. The aim of using three different datasets is to validate our proposed speech emotion recognition model cross-culturally. Features fusion helps in gathering complementary information and boosting the accuracy of recognition.

Major contribution The emotion change detection (ECD) component is introduced along with the emotion assessment module. ECD helps in localizing the emotion change points in continuous speech signals and triggering an emotion recognition algorithm for accurate emotion classification. Most of the studies in the field of SER have used equal size segments for emotion assessment while in real life emotions persist for different periods. Emotion change detection marks the boundaries between different emotions based on how long they persist. These boundaries further help in speech segmentation for emotion assessment. The glottal features are used for emotion change detection. ECD module has considerably reduced the processing time and computational resources required to implement the proposed algorithm, instead of always an ON speech emotion recognition system. As most of the part of

continuous speech is neutral only the segments of speech where emotion change is observed are used for emotion assessment.

Implemented emotion assessment module using fused features set. The feature vector used for classification is composed of handcrafted features pitch, spectral, and prosodic derived from the raw audio signal and deep learned derived from spectrogram images. Feature fusion has boosted the accuracy of emotion assessment due to complementary information generated from different representations of speech signals.

The flow of the proposed work is as follows: In Section 2, we review the related works. In Section 3, we discuss the proposed methodology. In Section 4 there are Results and discussions and Section 5 contains the conclusion and future work.

2 Related work

After reviewing the work done in the field of speech emotion recognition it is found that lots of variation exist in terms of approaches, emotion representation model, unit of analysis, choice of features, feature selection algorithms, and classifiers exist [24]. Two different models discrete and in continuous dimensions were used for emotion representation. In the discrete model, emotions are categorized into six basic emotions such as neutral, anger, disgust, fear, happiness, and sadness. In the continuous dimension, emotions are represented using two-axis arousal or activation and valence or positivity. Emotion recognition using a conventional pipeline requires feature extraction, feature selection, and emotion classification. A variety of features such as glottal, spectral, prosodic, voice quality features, and Teager energy operator were used for speech emotion recognition [13]. During heightened emotion change in vibration of the vocal fold is observed which further leads to variation glottal flow. Glottal features are used to capture the characteristics of the sound source and glottal flow and are among the prominent features used for speech emotion recognition. Glottal features are independent of language and robust against noise.

Nooteboom [23] Prosodic feature refers to the variations in melody, intonation, pauses, stresses, intensity, vocal quality, and accents of speech these features are based on the perception of humans. Utterance level prosodic features are widely used for speech emotion recognition [11]. Spectral features are extracted from the frequency domain and represent the characteristics of the vocal chard. Mel-frequency cepstrum coefficient (MFCC) features represent the spectral property of the speech signal. Recent studies have shown that spectral features also contain rich information about expressivity and emotion. The popularity of spectral features lies in the fact that vocal chord characteristics could be observed using spectral features [26]. Voice quality features used for SER are jitter, shimmer, and harmonics-to-noise ratio. Voice quality features provide insight into variations occurring in vocal tract characteristics during heightened emotions. Teager features are widely used for stress recognition these features help in detecting stresses occurring in muscles of the vocal tract due to changes in mental state. Spectral features provide frequency-domain metrics on your data. Both local and global features were used in emotion classification. The features are extracted from frames are called local features while global features are extracted from the complete utterance. Local features are good in capturing local dynamic information of emotion in speech. The global features take statistics of the features in a whole utterance for speech emotion classification. Features used for speech emotion recognition were extracted from frame, word, or a complete sentence. Feature

extraction is followed by feature selection in this step features that are highly correlated or redundant are removed.

Different algorithms were used for feature selection such as principal component analysis (PCA), Linear discriminant analysis (LDA), Extreme learning machines, swarm intelligence, etc. The classification algorithm used for SER is linear regression, K-nearest neighbor (KNN), support vector machines(SVM), Gaussian mixture model (GMM), Hidden Markov model, Artificial neural network, etc. A lot of time, effort, and domain knowledge is required to generate an optimal feature set still there is no accepted feature set that can distinguish different emotions.

Currently, deep learning algorithm has replaced the conventional method for speech emotion recognition as they are capable of automatic generation of optimal features from training data without human intervention. Deep learning algorithms have shown good performance in several tasks such as image classification, object recognition, speaker recognition, speech and handwriting recognition, etc. Different variants of deep learning architecture such as Convolution neural network, long short term memory network, autoencoders, adversarial neural network, and RNN were employed for speech emotion recognition and have shown marked improvement in accuracy and performance compared to conventional approaches used for speech emotion recognition. Several researchers have recently used speech spectrograms or sub-bands of spectrograms for speech emotion recognition using CNN. As a result, it's reduced to an image classification problem. (Zhang et al.,2015) proposed CNN with different configurations with different features for speech emotion recognition. The prosody features usually focused on fundamental frequency (F0), speaking rate, duration, and intensity are not able to confidently differentiate angry and happy emotions from each other [18] proposed deep convolutional neural network (CNN) for speech emotion recognition using features extracted from spectrograms. The rectangular kernels of varied sizes were proposed to learn discriminative from the spectrogram. The proposed method was evaluated on Emo-DB and Korean datasets and has shown better performance in comparison to other state-of-the-art techniques proposed for SER. Lu et al. [16] proposed long short-term memory (LSTM) and convolutional neural network (CNN) for SER. LSTM extracts the temporal context features of the speech signals and CNN extracts high-level emotional features from low-level features for emotional classification. The proposed method obtained accuracies of 49.15%, 85.38%, and 37.90% on eNTRAFACE'05, RML, and AFEW6.0 databases.

Zhang et al. [29] proposed integrated distributed-gender feature and gender-driven feature with spectrogram features for speech emotion recognition using convolutional neural network (CNN) and bi-directional long short-term memory (BLSTM) and reduced recognition error rate to 45.74%. Zhang et al. [30] proposed low-level descriptors and spectrographic features for speech emotion recognition using CNN and reduced the error rate to 36.91%. Guo et al. [8] used amplitude and phase information for speech emotion recognition and recognition error to 33% the dataset used to experiment are EMO-DB). Hajarolasvadi and Demirel [9] proposed 3D CNN for speech emotion recognition using Mel Frequency Cepstral Coefficients (MFCC), pitch, and intensity feature and features extracted from spectrogram derived from same frame dataset used for experimentations were Surrey Audio-Visual Expressed Emotion (SAVEE), Ryerson Multimedia Laboratory (RML), and eNTERFACE'05 databases and obtained results superior to state-of-the-art methods reported in the literature [19]. A proposed novel architecture of ADRNN networks for speech emotion recognition. The local correlations feature and global contextual information learned from Mel-spectrogram are passed as input to BI LSTM for speech emotion recognition. Accuracies of 90.78% and 85.39% and unweighted accuracy

of 74.96% and 69.32% were obtained on speaker-dependent and speaker-independent using Berlin EMODB and IEMOCAP.datasets. Badshah et al. [2] proposed a convolution neural network using affect-salient features for SER.

Anvarjon et al. [1] proposed frequency features extracted from the speech data to predict emotions. The accuracies of 92.02% and 77.01% are obtained on the IEMOCAP and Emo-DB datasets, respectively. Mustaqeem et al. [21] proposed a computationally efficient method for SER. Instead of using entire utterance only key segments of utterance were proposed for speech emotion analysis using CNN. The proposed method achieved accuracies of 72.25%, 85.57%, and 77.02% on IEMOCAP, EMO-DB, and RAVDESS datasets, respectively. [20] Proposed a one-dimensional dilated convolutional neural network (DCNN) to learn spatial salient emotional features and learn long-term contextual dependencies from the speech signals. [27] Proposed low-level descriptors (LLD), segment-level features extracted from speech Mel-spectrograms, and features extracted from the complete utterance. In the proposed speech emotion recognition model three classifiers deep neural network (DNN), a convolution neural network (CNN), and a recurrent neural network (RNN) were integrated. The integrated model achieved weighted accuracy of 57.1% and unweighted accuracy of 58.3% on the IEMOCAP dataset. The accuracy of the proposed system was significantly better in comparison to the individual classifier. Zhang et al. [28] proposed multi- CNN to learn multimodal audio features from spontaneous speech for emotion recognition. Through a fusion of different features, complementary information could be generated which significantly improved the accuracy of emotion recognition. The proposed model was evaluated on AFEW5.0 and BAUM-1 s datasets. Motivated by the well-established success of pitch, and spectral features for SER, as well as by the recent popularity of spectrogram images, we propose a new SER approach by concatenating pitch and spectral features with the spectrogram image features extracted by a CNN.

2.1 Preliminaries

2.1.1 Convolution Neural Network

The convolutional neural network is the most popular and established deep learning algorithm. CNN has shown exceptionally good performance in several tasks related to a variety of domains with benchmark performance in image classification. CNN mimics the structure of the human brain. CNN takes input such as image and systematically apply filters to input data to learn discriminatory features. The filters applied to input help to learn abstract concepts such as boundaries and edges, etc. Although filters could be handcrafted the advantage of using CNN is that it automatically learns the filters during the training process in the context of the problem at hand. CNN requires lesser pre-processing in image classification when compared to other classification algorithms. CNN automatically extracts features from the input which was earlier performed by a human in traditional algorithms. The major advantage of CNN there is no need for prior knowledge for feature engineering and they are highly scalable to massive data sets. The building blocks of CNN are convolution layers and fully connected layers. A typical CNN architecture is composed of multiple convolutions and pooling layers stacked over each other followed by one or more fully connected layers. Two important operations of CNN are convolution and pooling. The convolution operation is the application of filters to input such as an

image that results in inactivation. The repeated application of filters on input results in an activation map indicating the strength and location of the detecting feature in an Input. Pooling operation reduces the number of features by extracting dominant features which further reduces the processing time and computational complexity of algorithms. The fully connected layer contains a feature vector which is further used for classification.

2.1.2 Spectrograms

Spectrograms are a time-frequency visual representation of a signal produced by a short-time Fourier transform (STFT) [2]. Spectrograms carry both vocal tract and excitation signals, as well as amplitude and phase information. They are critical to analyzing speech signals both in the time and frequency domain. Although there are other representations of speech signal such as wavelet transform and Wigner–Ville distribution spectrograms offer distinct advantages and are preferred in practical applications. As a result of increasingly complex modeling approaches and large datasets, their value is growing, even if more stringent feature extraction could potentially worsen overall performance.

The spectrogram is the image of sound representing changing frequency content of the signal concerning time. The frequency of the signal is represented on the y-axis, and time on the x-axis in the 2D spectrogram. The lowest frequencies are represented at the bottom of the spectrogram while the higher frequencies are represented at the top. The colors are used to represent the amount of energy in the signal. Higher energy regions in spectrogram caused by events such as vocal fold closures, formants, and harmonics are shown using darker color; lighter color such as white is used to represent the region of little energy such as silence. Spectrograms are of two types' narrowband and wideband spectrogram. A Narrowband spectrogram is used to show the characteristics of the source like the vibration of vocal folds while a wideband spectrogram is used to investigate the characteristics of the vocal tract such as vocal tract resonance (formants).To generate a spectrogram of sound signal; the signal is first divided into smaller overlapping segments called frames followed by short-Fourier transform. Figure 1 shows a spectrogram for emotion anger stress and neutral.

$$X(w,m) = STFT(x(n)) = \sum_{n=0}^{R-1} x(n-m)w(n)e^{-jwn} \quad (1)$$

Where x(n) denotes the input signal at time n and w(n) denotes the R-length window function.

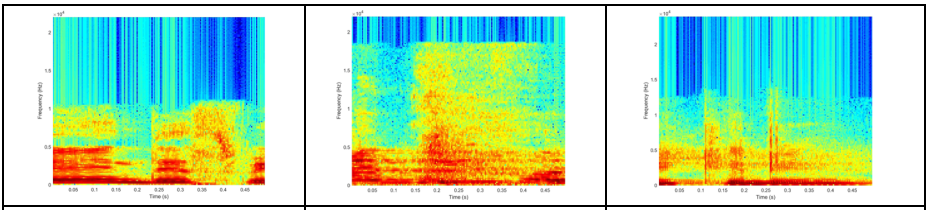

Fig. 1 Spectrogram of emotion anger, stress, and neutral

3 Proposed methodology

The proposed methodology for emotion recognition (anger, stress, and neutral) consist of three important steps pre-processing, feature extraction, and classification and the block diagram in Fig. 2 showing the details of each step used in the proposed methodology for anger and stress recognition.

3.1 Pre-processing

To implement the proposed work three publically available datasets Toronto Emotional Speech dataset (TESS), Berlin Emotional Database (EmoDB), and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) were downloaded. The details of each dataset are available in Section 4.1. After downloading the dataset corpus related to emotion anger, stress, and neutral were extracted and stored separately. Then, each speech signal is segmented into shorter overlapping frames of equal length. Each frame has a 50% overlap with the previous one. This step results in the division of each speech signal into n frames. These frames are further analyzed using glottal features to identify the frame which shows glottal asymmetry. The frames of speech where glottal asymmetry was observed were further used for emotion assessment.

3.2 Glottal asymmetry detection for emotion change analysis

During normal phonation both vocal folds left and right always show symmetric oscillation. The oscillatory pattern of the vocal folds gets asymmetric due to the influence of emotion or pathological conditions. Glottal symmetry is an important feature used in determining emotional state and abnormality in voice due to pathological conditions. For estimation of glottal signal, we have used inverse filtering. The glottal signal could be obtained using Eq. 2, where

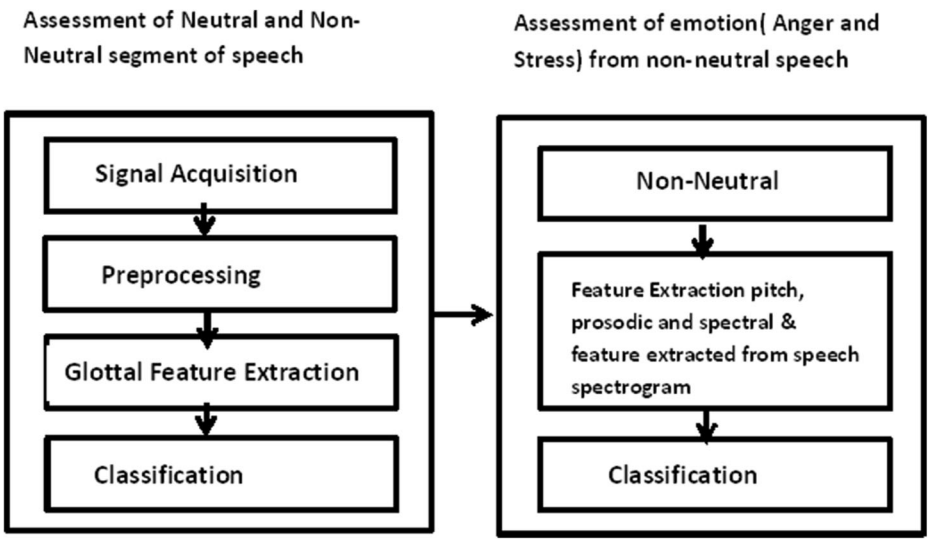

Fig. 2 Methodology for assessment of emotion anger and stress

G (z) represents glottal modal, V(z) represents vocal tract (VT) system function, and R(z) is the effect of radiations from the lip.

$$G(z) = \frac{S(z)}{V(z)R(z)} \tag{2}$$

The shape of the glottal pulse is composed of To is the pulse opening phase, Tc is the pulse closing phase, Uo represents peak volume velocity of glottal pulse which occurs at tp, and FG signifies glottal frequency of oscillation. To opening phase of the pulse and Tc closing phase of the pulse could be determined using an equation. The T_c and T_0 is calculated using Eqs. 3 and 4 respectively.

$$T_c = \left(\frac{1}{FG}\right)\left[\frac{cos^{-1}[(k-1)/k]}{2\pi}\right] \tag{3}$$

$$T_o = \frac{1}{FG} - \mathrm{Tc} \tag{4}$$

After determining T_o and T_c the glottal symmetry of the pulse is determined using Eq. 5. The glottal symmetry is defined as the ratio of the closing phase and opening phase.

$$GS = \frac{\mathrm{Tc}}{\mathrm{To}} \tag{5}$$

Several samples of neutral speech were used to extract T_c and T_o value related neutral speech. The recorded T_c and T_o values are used to calculate glottal symmetry related to each recorded sample. The mean value of glottal symmetry is obtained from the collected data. The control chart is further used in identifying part speech where glottal asymmetry is observed. The control chart shown in Fig. 3 is the diagram used to monitor the variations in particular characteristics of the process over time. After analyzing the sufficient points mean value for the process is calculated. The mean value is used to calculate the upper and lower control limits. For the process to be stable values must lie between the upper or lower control limit. The value which lies above or below the upper or lower control shows process instability. Adding (3 x σ

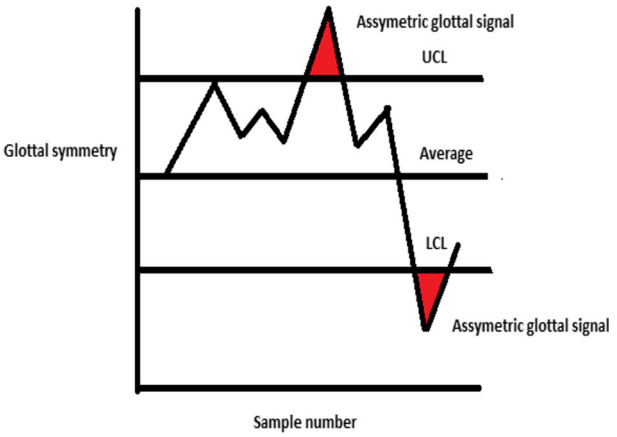

Fig. 3 Description of implementing control charts for emotion change detection

to the average) for the UCL and subtracting (3 x σ from the average) for the LCL where σ is used for standard deviation.

If the values of glottal symmetry obtained at the current position go beyond the upper or lower control limit of reference neutral speech emotion assessment events are triggered.

3.3 Feature extraction

In the proposed approach two sets of features were used for the assessment of emotion anger, stress, and neutral. These features are derived from the frame which shows glottal asymmetry.

3.3.1 Handcrafted features

The first feature vector is composed of 11 handcrafted features RMS, ZCR, Spectral centroid, Spectral Entropy, Spectral roll-off, mean pitch, max pitch, min pitch, tempo, low energy, spectral irregularity were extracted from each frame of speech showing glottal asymmetry. The RMS, ZCR, Spectral centroid, Spectral Entropy, and Spectral roll-off are derived using Eqs. 6,7,8,9, and 10 respectively.

- **RMS** is the measurement of energy in the signal.

$$\mathrm{ZCR} = \frac{1}{\mathrm{M}-1}\sum_{\mathrm{m}=0}^{\mathrm{M}-1}|\mathrm{sign}(\mathrm{x}(\mathrm{m})) - \mathrm{sign}(\mathrm{x}(\mathrm{m}-1))| \tag{6}$$

- **ZCR** The zero-crossing rate indicates the frequency at which the signal crosses the zero amplitude level.

$$\mathrm{ZCR} = \frac{1}{\mathrm{M}-1}\sum_{\mathrm{m}=0}^{\mathrm{M}-1}|\mathrm{sign}(\mathrm{x}(\mathrm{m})) - \mathrm{sign}(\mathrm{x}(\mathrm{m}-1))| \tag{7}$$

M = total number of samples in processing window, X(m) = is the value of mth sample.

- The spectral centroid indicates at which frequency the energy of a spectrum is centered. here S(k) is the spectral magnitude at frequency bin k, f(k) is the frequency at bin k.

$$fc = \frac{\sum_k S(k)f(k)}{\sum_k S(k)} \tag{8}$$

- **Spectral Entropy** measures signals irregularity.

$$\mathrm{SE}\left[\mathrm{f}_1,\mathrm{f}_2\right] = -\frac{1}{\log\left[\mathrm{N}\left[\mathrm{f}_1,\mathrm{f}_2\right]\right]}\sum_{\mathrm{f}_{\mathrm{i}=\mathrm{f}_\mathrm{k}}}^{\mathrm{f2}}\mathrm{P}_\mathrm{n}(\mathrm{f}_\mathrm{i})\log(\mathrm{P}_\mathrm{n}(\mathrm{f}_\mathrm{i})) \tag{9}$$

Where P n (f i) represents the probability of the ith frequency component, SE corresponding to the frequency range [f 1, f 2].

- **Spectral roll-off** represents the frequency below which total spectral energy is concentrated.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \tag{10}$$

Where Mt[n] is the magnitude of the Fourier transform at frame t and frequency bin n.

- **Pitch statistic** mean pitch, max pitch, min pitch.
- **Tempo** represents how fast or slow sound is.
- **Low energy** measures the number of frames whose RMS energy is less than the threshold.

3.3.2 Automatically extracted features

The second set of features was automatically extracted from the spectrogram derived from speech frames where glottal asymmetry was observed. The features were extracted from the spectrogram using the proposed CNN architecture shown in Fig. 4. The spectrogram is a visual representation of STFT where the horizontal axis represents the time and the vertical axis represents the frequency of the signal in that short frame. In a spectrogram, at a particular time point and a particular frequency, dark colors illustrate the frequency in a low magnitude, whereas light colors show the frequency in higher magnitudes. Spectrograms are perfectly suitable for a variety of speech analyses including SER [18].

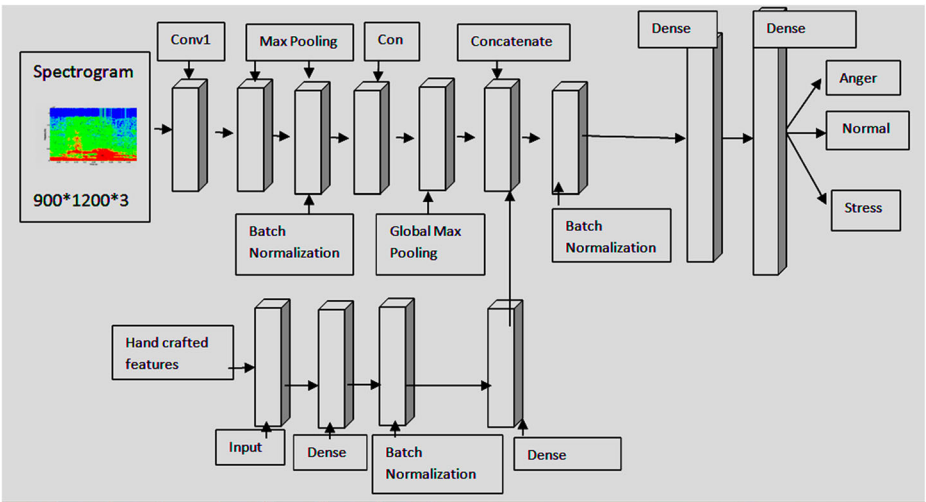

Fig. 4 Description of the proposed convolutional neural network architecture used for emotion assessment. Two sets of features are used for the assessment of emotions (anger, stress, and neutral) first set of features derived from speech spectrogram using CNN and the second set of features (handcrafted features) were derived from speech frame

3.4 Classification

For emotion classification, both sets of features were concatenated followed by batch normalization. The batch normalization layer performs the standardization and normalization on the input coming from the previous layer. Batch normalization smoothens the loss function that in turn optimizes the model parameters and speeds up the training process. Which are further passed to a fully connected layer. The final fully connected layer provides voting of each emotion classes.

4 Results and discussions

4.1 Dataset

4.1.1 Toronto Emotional Speech Set (TESS) dataset

TESS [5] is an emotional speech dataset of seven emotions (anger, happiness, pleasure, fear, disgust, surprise, neutral, and sadness). Dataset was recorded in English and recording consisted of 200 words by two actresses of ages 26 and 64 years. An audiometric test was carried out to keep the thresholds within the normal range.

4.1.2 EMO-DB dataset

Burkhardt et al. [3] Berlin Emotional Database (EmoDB) is a publically available emotional speech database containing utterances by 10 German obtained from Emergency call centers consisting of seven emotions (i.e., anger, fear, disgust, and boredom, happy, neutral, and sad). Background noise is present in the recording and recording is made by non-actors. The numbers of utterances in the dataset are 500.

4.1.3 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [14]

This is emotional speech dataset in the English language consists of 8 emotions (happy, disgust, anger, surprise, calm, fear, sad and neutral, and sad). The 24 actors consisting of 12 males and 12 females have used to record the 8 emotions. The total recorded utterances are 1440 wav files at a sampling rate of 48,000 Hz.

4.2 Performance evaluation metrics

The evaluation metric used to evaluate the performance of the proposed model are confusion matrix, categorical accuracy, and cross-entropy loss.

4.2.1 Confusion matrix

The detailed data samples correctly and incorrectly classified can be shown in terms of the N*N matrix called as confusion matrix where the number of classes denoted by N. The number of data samples that are classified correctly forms the diagonal elements in the

confusion matrix while off-diagonal elements in confusion matrix depict data samples that are misclassified. So higher is the classification accuracy if diagonal values are higher.

4.2.2 Categorical accuracy

This is calculated by dividing correctly classified samples by a total number of samples. Accuracy can be training accuracy or validation accuracy. The accuracy obtained on the training set is called training accuracy and the accuracy obtained on the validation set is called validation accuracy.

4.2.3 Cross entropy loss

The diversion of actual prediction from a true prediction by classification model is measured as cross-entropy loss whose probability values lie between 0 and 1. Cross entropy loss is used for optimizing a deep learning model.

4.3 Experimental setup for emotion assessment

Three experiments were carried out to evaluate the proposed architecture. The peltarion platform was used to perform experiments. The platform is GUI-based and could be used for developing, deploying, and managing deep learning models. It is a cloud-based platform capable of advanced modeling, data access, data pre-processing, training deep learning models, evaluating the model, and visualizing results. To perform the experiment dataset is uploaded on the platform followed by model architecture design, hyper-parameter setting, model training, and model evaluation. The convolutional neural network model shown in Fig. 5 was designed, implemented, and evaluated using the peltarion platform.

Experiment-1 was carried out using Toronto Emotional Speech Set (TESS) dataset. The confusion matrix obtained after implementing the proposed model on the TESS dataset is shown in Fig. 6. The classification accuracy obtained for class anger is 90.9% with 6.1% and 3% misclassified as normal and stress. The classification accuracy obtained for class normal is 96.8% with 3.2% misclassified as stress. The classification accuracy for class stress is 100%. So the highest accuracy of classification shown by the model is for class stress and lowest for anger. The overall training accuracy of the model in predicting all three emotions is 0.937(93.7%), Macro Precision 0.937(93.7%), Macro Recall 0.937(93.7%), Macro F1-Score 0.937(93.7%), and Cross-Entropy Loss 0.199. The overall validation accuracy of the model in predicting all three emotions is 0.959 (95.9%), Macro Precision 0.955 (95.5%), Macro Recall 0.959 (95.9%), Macro F1-Score 0.956 (95.6%), and Cross-Entropy Loss 0.272.The results of experiment-1 are summarized in Table 1. Figure 7 shows the training and validation cross-entropy loss plot for Experiment-1 on the TESS database and Fig. 8 shows the training and validation categorical accuracy plot for Experiment-1 on the TESS database.

Experiment-2 was carried out using the EMO-DB dataset. The confusion matrix obtained using the proposed model is shown in Fig. 9. The classification accuracy for class anger is 93.9% with 6.1% of data samples misclassified as normal. The classification accuracy for class normal is 93.5% with 6.5% of data samples misclassified as anger. The classification accuracy for class stress is 100%. So the highest classification accuracy is obtained for class stress and lowest for class anger. The overall training accuracy of the model in predicting all three emotions is 0.956(95.6%), Macro Precision 0.958(95.8%), Macro Recall 0.958(95.8%), Macro

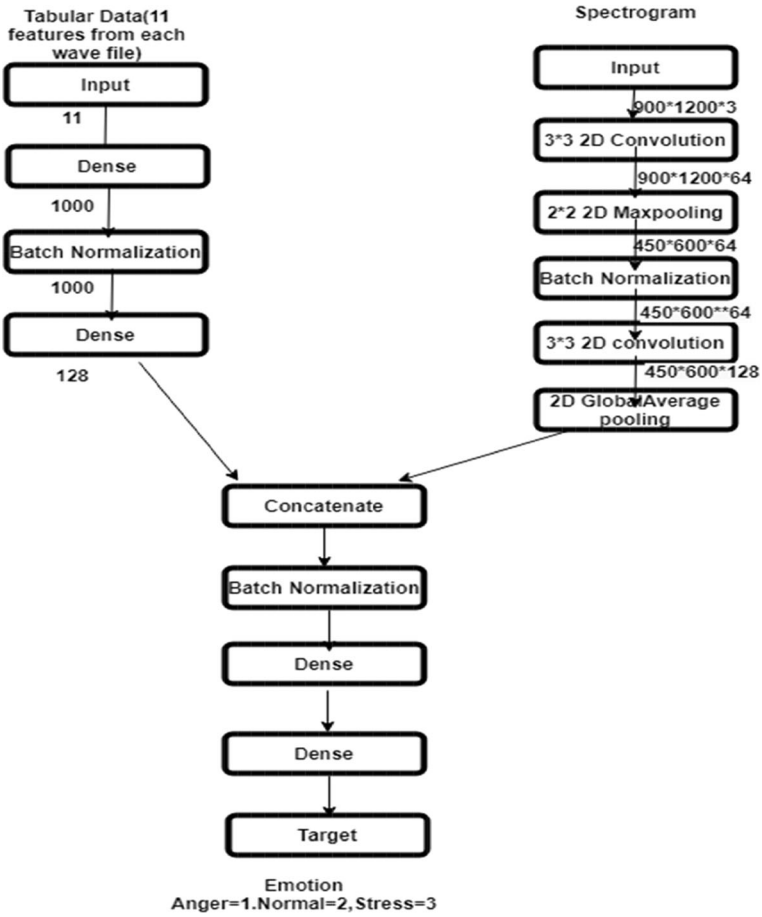

Fig. 5 Description CNN architecture used for speech emotion recognition with two different sets of inputs (1) handcrafted features (2) Spectrogram Images. Emotions detected are anger, stress, and neutral

F1-Score 0.958(95.8%), and Cross-Entropy Loss 0.094. The overall validation accuracy of the model in predicting all three emotions is 0.975(97.5%), Macro Precision 0.975(97.5%), Macro Recall 0.974(97.4%), Macro F1-Score 0.974(97.4%), and Cross-Entropy Loss 0.069. The results of experiment-2 are summarized in Table 2. The Fig. 10. Show training and validation cross-entropy loss plot for Experiment-2 on Emo-DB and Fig. 11. Show training and validation categorical accuracy plot for Experiment-2 on Emo-DB database.

Table 1 The results of experiment-1

Evaluation Metric	Training (T)	Validation(V)
Cross Entropy Loss	0.199	0.272
Categorical Accuracy	0.937	0.959
Macro Precision	0.937	0.955
Macro Recall	0.937	0.959
Macro F1-Score	0.937	0.956

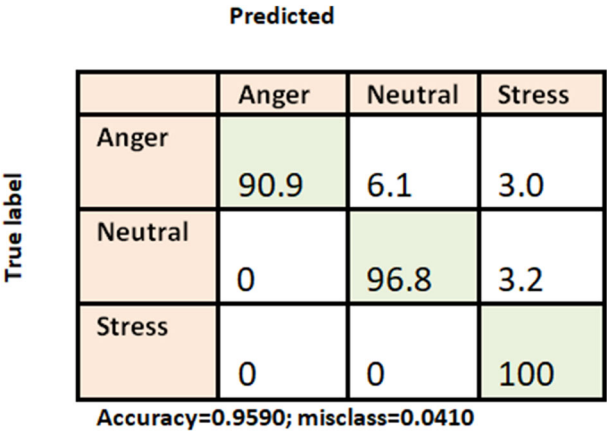

Fig. 6 Confusion Matrix for Experiment-1 using TESS database

Experiment-3 was carried using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Figure 12 is a confusion matrix of the experiment performed on the RAVDESS dataset. The classification accuracy of class anger is 100%, 90.3% of data belonging to class normal are correctly classified as normal and 9.7% are misclassified as anger. All the data samples of class stress are correctly classified showing an accuracy of 100%. So the highest accuracy of classification by the model is shown for class anger and stress, and lowest for normal. The overall training accuracy of the model in predicting all three emotions is 0.967 (96.7%), Macro Precision 0.972 (97.2%), Macro Recall 0.968(96.8%), Macro F1-Score 0.96.9(96.9%), and Cross-Entropy Loss 0.081. The overall validation accuracy of the model in predicting all three emotions is 0.967(96.7%), Macro Precision 0.967(96.7%), Macro Recall 0.967(96.7%), Macro F1-Score 0.974(97.4%), and Cross-

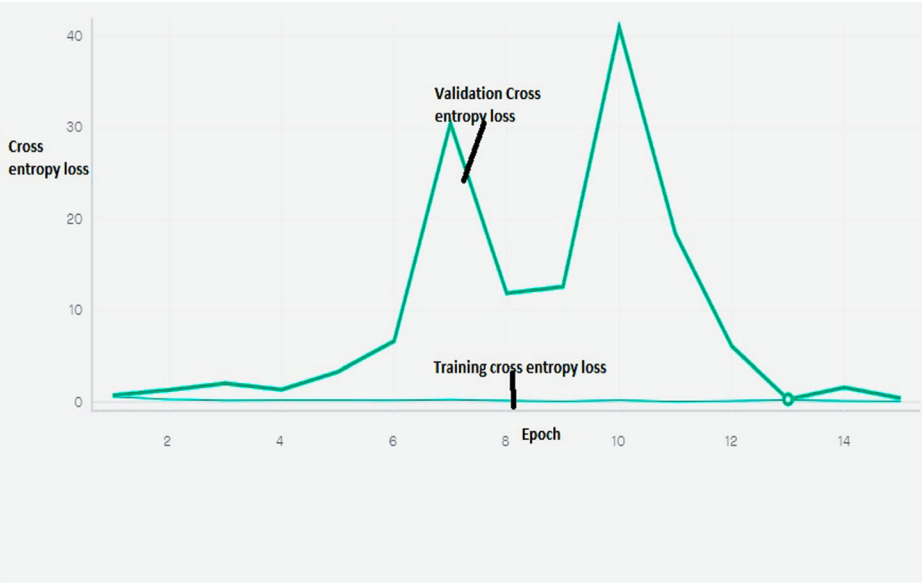

Fig. 7 Show training and validation cross-entropy loss plot for Experiment-1 on TESS database

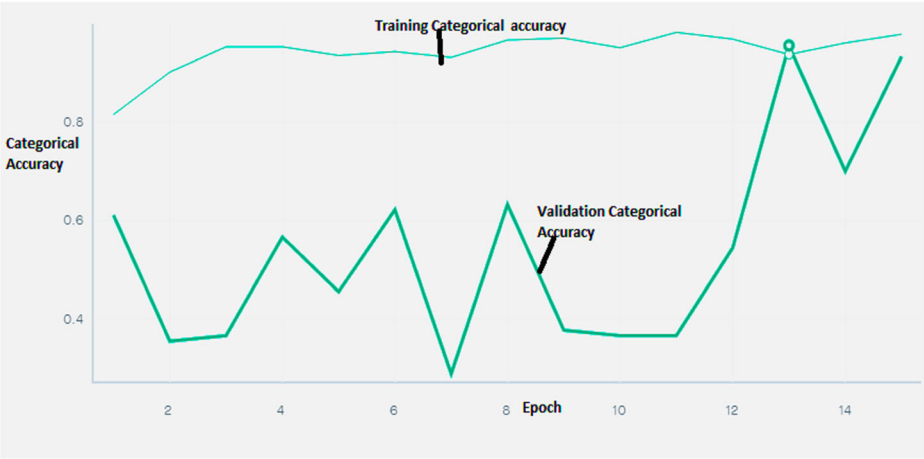

Fig. 8 Show training and validation categorical accuracy plot for Experiment-1 on TESS database

Table 2 Results of Experiment-2

Evaluation Metric	Validation	Training
Cross Entropy Loss	0.094	0.069
Categorical Accuracy	0.956	0.975
Macro Precision	0.958	0.975
Macro Recall	0.958	0.974
Macro F1-Score	0.958	0.974

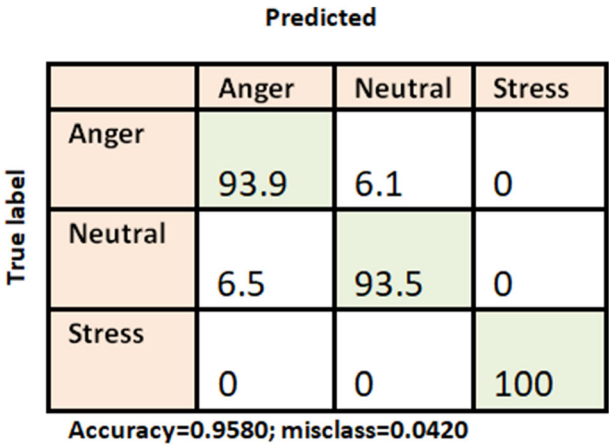

Fig. 9 Confusion Matrix for Experiment-2 using Emo-DB database

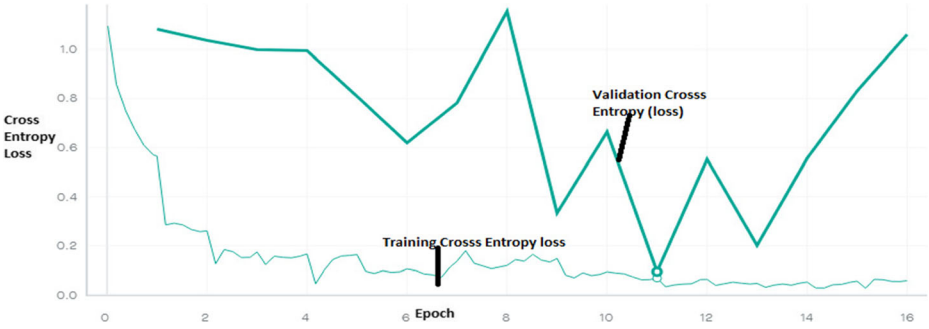

Fig. 10 Plot showing training and validation cross-entropy loss for Experiment-2 on Emo-DB database

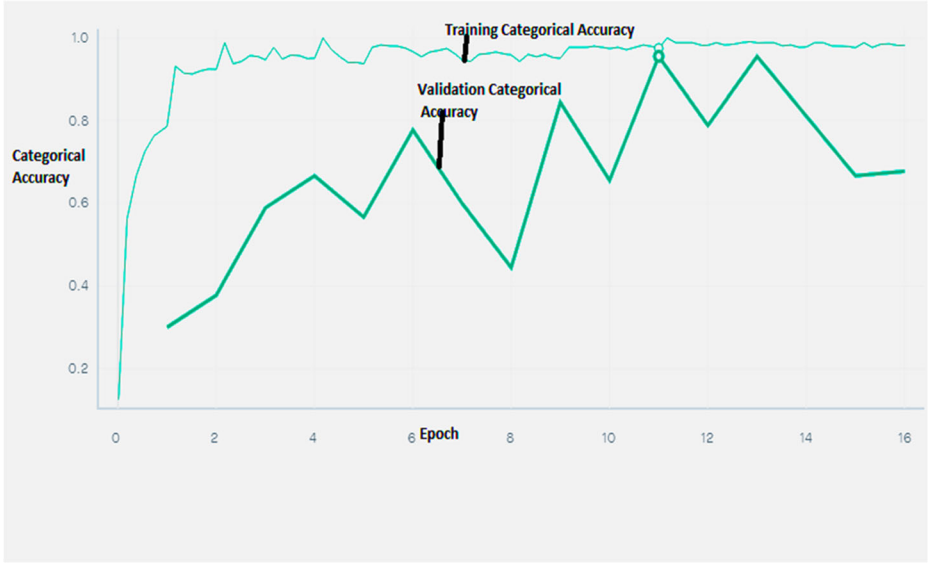

Fig. 11 Plot showing training and validation categorical accuracy for Experiment-2 on Emo-DB database

Entropy Loss 0.085. The results of experiment-3 are summarized in Table 3. The Fig. 13. Shows training and validation cross-entropy loss plot for Experiment-3 on RAVDESS database and Fig. 14. Shows training and validation categorical accuracy plot for Experiment-3 on RAVDESS database.

4.4 Training the proposed CNN

The proposed model was trained on 600 speech files belonging to each dataset. 200 files are belonging to each category (anger, stress, and neutral). From each speech file features proposed in the section were extracted and each file is also transformed into spectrograms. To evaluate the performance of the proposed architecture on each dataset, datasets are first to split into training and validation sets. 80% of data is used for training while 20% of data is used for validation. The total number of epochs used for training the network is 100; batch size 64,

Table 3 Results of Experiment-3

Evaluation Metric	Validation	Training
Cross Entropy Loss	0.081	0.085
Categorical Accuracy	0.967	0.967
Macro Precision	0.972	0.967
Macro Recall	0.968	0.967
Macro F1-Score	0.969	0.974

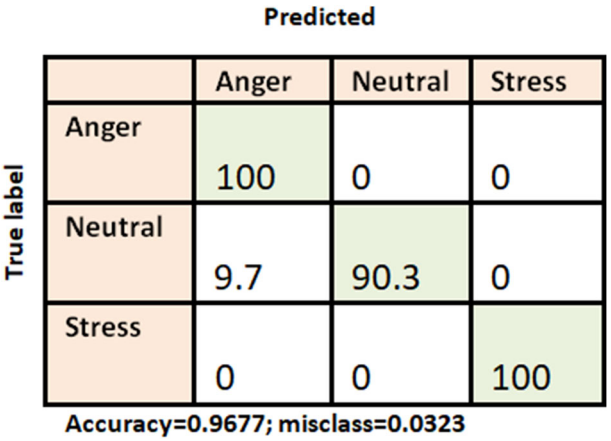

Fig. 12 Confusion Matrix for Experiment-3 using RAVDESS database

the initial learning rate is set to 0.00025, and Optimizer used is Adam. The exponential decay rate $\beta 1 = 0.9$ and $\beta 2 = 0.999$. Hyper-parameters values are initialized based on a heuristic. The softmax activation function is used at the output layer and all the other layers use the ReLU activation function. The probability associated with the output is determined using the softmax function. The summary of hyper-parameter settings is Table 4.

4.5 Performance comparison of proposed method

The performance of the proposed method is further compared with the state of art techniques used for speech emotion recognition (Table 5).

5 Conclusion and future work

We are always in search of methods to improve the performance of SER systems. The traditional methods used for speech emotion detection used pipeline architecture where each module has to be optimized separately. In the proposed work we have used end to end approach by replacing pipeline architecture with a single convolution neural network. The system extracts useful features from spectrogram images along with that we have also passed handcrafted features such as pitch, spectral, and energy-related features to gather

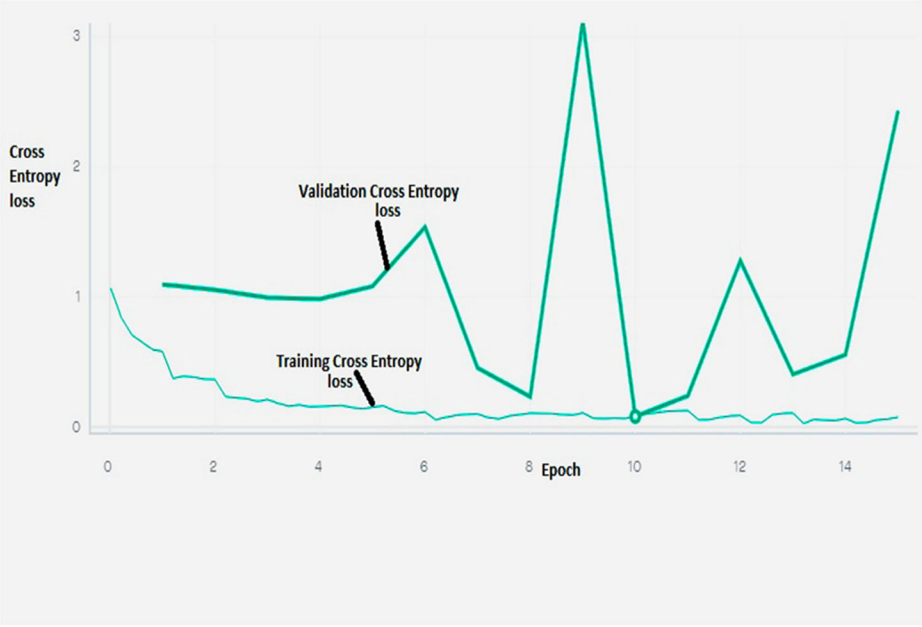

Fig. 13 Plot showing training and validation cross entropy loss for Experiment-3 on RAVDESS database

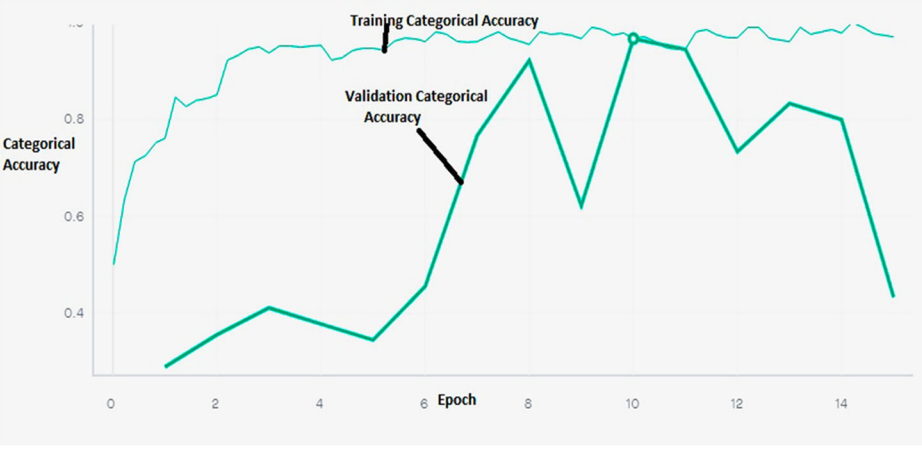

Fig. 14 Plot showing training and validation categorical accuracy for Experiment-3 on RAVDESS database

complementary information. In the proposed work we have introduced an emotion change detection module before emotion assessment has considerably reduced the computational complexity as well as boosted the performance of the proposed algorithm by processing only those segments of speech that differ from neutral speech patterns. A similar work done earlier has used pretrained networks for emotion classification. They are heavy architecture requiring more memory and more processing time. So we have proposed our own CNN architecture with fewer layers and feature fusion. The proposed research provides insight to researchers

Table 4 Hyper-parameter setting

Hyperparameter	Value
Initial Learning rate	0.00025
Number of Epochs	100
Batch Size	64
Learning Rate decay	0
$\beta_{1\ rate}$	0.9
$\beta_{2\ rate}$	0.999
Data set Split Training/Validation set	80% / 20%
Optimizer	Adam
No of Filters used in 1st Convolutional Layer	64
No of Filters used in 2nd Convolutional Layer	128

Table 5 Shows the comparison of the proposed method with other well established methods used for emotion detection through speech

Author	The approach used for SER	Accuracy Achieved on different datasets
H. Meng et al. [8]	Proposed novel architecture attention-based convolutional recurrent neural networks(ADRNN) for SER dilated CNN with residual block and BiLSTM based on the attention mechanism	Obtained accuracies of 90.78% and 85.39% on Berlin EMODB on speaker-dependent and obtained 74.96% unweighted accuracy in the speaker-dependent and the 69.32% unweighted accuracy in the speaker-independent experiment on IEMOCAP dataset
Zhong et al. [9]	Proposed lightweight model (Convolutional neural network) to predict the emotional class	achieves 71.72% and 90.1% of unweighted accuracy (UA) on the well-known corpora IEMOCAP and Emo-DB respectively.
Farooq et al. [19]	Proposed deep convolutional neural network (DCNN) for SER	Obtained accuracy of 95.10% for Emo-DB, 82.10% for SAVEE, 83.80% for IEMOCAP, and 81.30% for RAVDESS, for speaker-dependent SER experiments.
Our Work	Proposed convolutional neural network with fused features	Obtained training (T) and validation (V) categorical accuracy of T=93.7%, V=95.6% for TESS, T=97.5%, V=95.6% for EMO-DB and T=96.7%, V=96.7% for RAVDESS dataset.

regarding the relevance emotion change module in emotion assessment as well as how a hybrid feature set could boost the accuracy of the system. The proposed research also tries to highlight that lighter CNN architecture must be preferred above heavy pretrained architecture to reduce computational complexity. Methods with high accuracy requiring fewer resources and lesser processing time are always preferred. Response time is one of the very important parameters for the system used to measure affective such as those in health care for monitoring mental health, customer perception, product marketing, education, etc. The limitations of the proposed study are a selection of handcrafted features is solely based on their popularity and use in past studies. Other features can be further explored for emotion change detection and emotion assessment to further boost the accuracy of emotion recognition. Still, there is a lot of scope for further improvement, the contribution of different features for specific action has not been paid enough attention and is yet to be explored. The architecture of the CNN used in the

above study is self-developed which could be further optimized using different hyper-parameter settings or different numbers of layers, sizes of spectrogram images. Instead of spectrogram other time-frequency representations such as Mel-spectrograms, different frequency bands of spectrograms, and single frequency spectrograms for further analysis. Future work in this direction could be exploring different sets of handcrafted features. Features extracted from different representations of speech signal such as Mel-spectrograms, different bands of spectrograms could be fused to see how they affect the accuracy of the proposed method.

Authors' contributions The methodology was performed by Shalini Kapoor., Literature survey is jointly done by all authors.

Data availability Availability of data materials, and software applications which is developed by present authors.

Code availability Code availability did by Shalini Kapoor.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval All authors ensure ethical approval.

Consent to participate All authors have the consent to participate.

References

1. Anvarjon T, Mustaqeem, Kwon S (2020) Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. Sens (Switzerland) 20(18):1–16. https://doi.org/10.3390/s20185212
2. Badshah AM et al (2019) Deep features-based speech emotion recognition for smart affective services. Multimed Tools Appl 78(5):5571–5589. https://doi.org/10.1007/s11042-017-5292-7
3. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: 9th European Conference on Speech Communication and Technology, vol 5, pp 1517–1520. https://doi.org/10.21437/interspeech.2005-446
4. Dhaka VS et al (2021) A survey of deep convolutional neural networks applied for prediction of plant leaf diseases. Sensors 21(14):4749. https://doi.org/10.3390/s21144749
5. Dupuis K, Pichora-Fuller MK (2010) Toronto Emotional Speech Set (TESS) | TSpace Repository. [Online]. Available: https://tspace.library.utoronto.ca/handle/1807/24487. Retrieved June 19, 2020
6. Fink G. (2016) Stress, Definitions, Mechanisms, and Effects Outlined: Lessons from Anxiety, in Stress: Concepts, Cognition, Emotion, and Behavior: Handbook of Stress. Elsevier, pp. 3–11. https://doi.org/10.1016/B978-0-12-800951-2.00001-7
7. Govoreanu VC, Neghina M (2020) Speech emotion recognition method using time-stretching in the preprocessing phase and artificial neural network classifiers. In: Proceedings – 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing, ICCP 2020, pp 69–74. https://doi.org/10.1109/ICCP51029.2020.9266265
8. Guo L, Wang L, Dang J, Zhang L, Guan H, Li X (2018) Speech emotion recognition by combining amplitude and phase information using convolutional neural network. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol 2018-Septe, pp 1611–1615. https://doi.org/10.21437/Interspeech.2018-2156
9. Hajarolasvadi N, Demirel H (2019) 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. Entropy 21(5):479. https://doi.org/10.3390/e21050479

10. Ijaz MF, Attique M, Son Y (2020) Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Sensors 20(10):2809
11. Jiang L, Tan P, Yang J, Liu X, Wang C (2021) Speech emotion recognition using emotion perception spectral feature. Concurr Comput Pract Exp 33(11):e5427
12. Juslin PN, Scherer KR (2008) Speech emotion analysis. Scholarpedia 3(10):4240
13. Kadiri SR, Alku P, Yegnanarayana B (2020) Comparison of glottal closure instants detection algorithms for emotional speech. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 7379–7383
14. 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. https://doi.org/10.1371/journal.pone.0196391
15. Low DM, Bentley KH, Ghosh SS (2020) Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investig Otolaryngol 5(1):96–116
16. Lu G, Yuan L, Yang W, Yan J, Li H (2018) Speech emotion recognition based on long short-term memory and convolutional neural networks. Nanjing Youdian Daxue Xuebao (Ziran Kexue Ban)/J Nanjing Univ Posts Telecommun (Natural Sci) 38(5):63–69. https://doi.org/10.14132/j.cnki.1673-5439.2018.05.009
17. Mandal M, Singh PK, Ijaz MF, Shafi J, Sarkar R (2021) A tri-stage wrapper-filter feature selection framework for disease classification. Sensors 21(16):5571. https://doi.org/10.3390/s21165571
18. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans Multimed 16(8):2203–2213. https://doi.org/10.1109/TMM.2014.2360798
19. 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. https://doi.org/10.1109/ACCESS.2019.2938007
20. Mustaqeem, Kwon S (2021) MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert Syst Appl 167:114177. https://doi.org/10.1016/j.eswa.2020.114177
21. Mustaqeem, Sajjad M, Kwon S (2020) Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access 8:79861–79875. https://doi.org/10.1109/ACCESS.2020.2990405
22. Nayak J et al (2021) Intelligent system for COVID-19 prognosis: A state-of-the-art survey. Appl Intell 51(5):2908–2938
23. Nooteboom S, others (1997) The prosody of speech: melody and rhythm. Handb phonetic Sci 5:640–673
24. Schuller BW (2018) Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Commun ACM 61(5):90–99
25. Srinivasu PN, SivaSai JG, Ijaz MF, Bhoi AK, Kim W, Kang JJ (2021) Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. Sensors 21(8):2852
26. Tavi L (2019) Classifying females' stressed and neutral voices using acoustic–phonetic analysis of vowels: an exploratory investigation with emergency calls. International Journal of Speech Technology 22(3):511–520. https://doi.org/10.1007/s10772-018-09574-6
27. Yao Z, Wang Z, Liu W, Liu Y, Pan J (2020) Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. Speech Commun 120:11–19. https://doi.org/10.1016/j.specom.2020.03.005
28. Zhang S, Tao X, Chuang Y, Zhao X (2021) Learning deep multimodal affective features for spontaneous speech emotion recognition. Speech Commun 127:73–81. https://doi.org/10.1016/j.specom.2020.12.009
29. Zhang L, Wang L, Dang J, Guo L, Guan H (2018) Convolutional neural network with spectrogram and perceptual features for speech emotion recognition. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol.11304 LNCS, pp 62–71. https://doi.org/10.1007/978-3-030-04212-7_6
30. Zhang L, Wang L, Dang J, Guo L, Yu Q (2018) Gender-aware CNN-BLSTM for speech emotion recognition. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 11139 LNCS, pp 782–790. https://doi.org/10.1007/978-3-030-01418-6_76