Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional Neural Network

Kishor Bhangale and K. Mohanaprasad

Abstract In recent years, speech emotion recognition (SER) has engrossed more attention in speech processing because of its potential in various speech-based intelligent systems. In deep learning algorithms to capture discriminative features of the audio emotion samples, a large number of features are required, which increases the computational complexity of the network. This paper presents a three-layered sequential deep convolutional neural network (DCNN) based on mel frequency log spectrogram (MFLS) for emotion recognition. Mel frequency log spectrogram that confines the salient information from the emotion speech corpus and two-dimensional DCNN. Exploratory outcomes on the Berlin Emo-DB dataset show that the proposed method gives 95.68 and 96.07% accuracy for the speaker-dependent and speaker-independent approaches. The performance of the proposed method is compared with CNN and CNN-LSTM on the Berlin Emo-DB dataset and results in improved accuracy.

Keywords Speech emotion recognition · Deep convolutional neural network · Mel frequency log spectrogram

1 Introduction

Speech emotion recognition (SER) is a crucial part of human–computer interaction (HCI) as speech is an efficient, fast and essential way of human interaction. This system eases the natural communication with the machine utilizing voice instructions rather than traditional input devices [1]. SER has widespread applications such as audio conferencing, interactive robot, call centre dialogue, aboard vehicle driving system, interactive game designing, medical psychological analysis, online learning and tutoring system [2, 3].

Determination of human speech emotion is an idiosyncratic task and can be used for any standard for any SER system. Human speech signal consists of verbal and para verbal information. The verbal information describes the meaning and context of

K. Bhangale · K. Mohanaprasad (✉)
SENSE, VIT, Chennai, India
e-mail: kmohanaprasad@vit.ac.in

A. Sivasubramanian et al. (eds.), *Futuristic Communication and Network Technologies*,
Lecture Notes in Electrical Engineering 792,
https://doi.org/10.1007/978-981-16-4625-6_24

the speech, whereas the para verbal information describes the tacit information such as the emotion expressed in the speech signal. The speech signal consists of different emotions like happiness, sadness, anger, fear, surprise, joy, disgust, boredom and neutral. The paralinguistic information is usually independent of the lexical content, language and speaker [4, 5].

Different emotion has an immense effect on the various characteristics of the speech signal such as short-term features like energy, pitch and format [6], long-term features like mean and standard deviation [7]; and prosodic features like pitch, speaking rate, intensity, voice variation and quality [8].

Traditional, machine learning (ML)-based SER systems have two major phases, such as feature extraction and classification. The performance of ML-based approaches is highly dependent on handcrafted features. In the past, many feature extraction techniques for SER system are implemented such as mel frequency cepstrum coefficients (MFCC) [9], principal component analysis (PCA), linear predictor coefficients (LPC), Gaussian mixture model (GMM), perceptual linear prediction coefficients (PLP) and hidden Markov model (HMM). The classification phase learns the extracted features and depicts the correct emotion. SER system used various classifications algorithms such as support vector machine (SVM), K-nearest neighbour classifier (KNN) and artificial neural network (ANN). The performance of the ML classifiers depends upon raw features, database size, professional knowledge and manual tuning of features which are labour expensive [10].

In recent years, deep learning (DL) emerges as the advanced field for SER, which represents the low-level speech features into the high-level hierarchical features. Jianwei Niu et al. presented DNN for the modelling of complex and nonlinear features of emotion speech training data. It resulted in 92.1% accuracy for the five-layered DNN with MFCC features [11]. Huang et al. presented CNN for the representation of salient hierarchical features in two stages. It achieved better accuracy for speaker-dependent approach [12]. Zheng et al. presented deep CNN (DCNN) for SER, which uses PCA for dimension reduction and interference suppression of the input log spectrogram. It resulted in an accuracy of 40% for IEMOCAP database [13]. Abdul Malik Badshah et al. presented an SER system based on DCNN with three fully connected layers that used spectrogram. It resulted in 84.3% accuracy of the Berlin Emotion database [14]. Jianfeng Zhao et al. presented 1D CNN-long short-term memory (1D-CNN-LSTM) and 2D CNN-LSTM (2D-CNN-LSTM) to discover local and global emotion-specific features. It resulted in 95.33 and 95.89% accuracy on the Berlin Emo-DB database for speaker-dependent (SD) and speaker-independent (SID) approaches [15].

DL algorithms have several advantages such as the capability to deal with complex speech structure and features; ability to deal with un-labelled data; no need for feature tuning; and ability to handle more massive datasets. Though SER has made tremendous progress still, it faces many challenges such as variability in individual, variability in environmental conditions and effect of noise, generalizing the model for the distinct dataset and recognition of subtle expression. Therefore, there is a need for the design of a robust SER system that can overcome these limitations.

This paper presents three-layered sequential deep convolutional neural network (DCNN) for the speech emotion recognition that accepts two-dimensional mel frequency log spectrogram (MFLS) as input. For the performance evaluation Emo-DB database, this consists of seven acted emotions samples such as anger, boredom, disgust, fear, neutral, happiness and sadness.

This paper is systematized as follows: Sect. 2 illustrates the proposed methodology, along with implementation details of MFLS and DCNN implementation in detail. Section 3 focuses on the discussion of simulation results. Finally, the conclusion and future perspectives are given in Sect. 4.

2 Proposed Methodology

The proposed SER based on a three-layered sequential DCNN consists of three CNN layers for emotion recognition (see Fig. 1). For the two-dimensional CNN mel frequency log spectrogram is given as the input which helps to capture salient features of the speech signal and minimization of the random noise present in the signal. Each CNN layer comprises three essential layers, such as the convolution layer (CL), exponential linear layer (ELU) and max pooling layer (MP). The fully connected (FC) layer followed after the MP layer of the third CNN and softmax classifier followed by a fully connected layer classifies the emotion speech signal.

2.1 Mel Frequency Log Spectrogram (MFLS)

The human emotion speech signal is one-dimensional. Thus to avail, the simplicity and advantages of the two-dimensional CNN, input emotion speech signal are converted into two-dimensional mel frequency logarithmic spectrum (see Fig. 2). Mel frequency gives the relation between the human ear and sound perception frequency [16]. The mel frequency scale (Mel) can be obtained from the linear frequency (f) using Eq. 1.

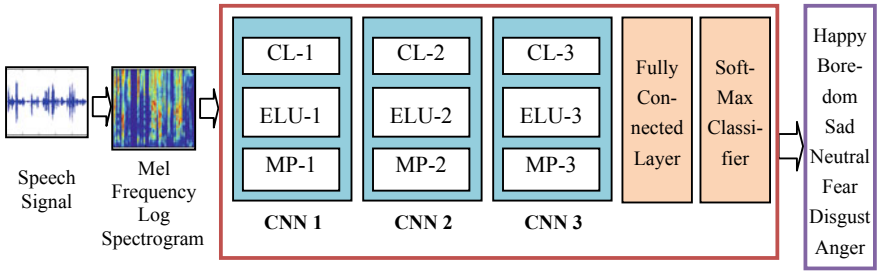

Fig. 1 Detailed process architecture of the proposed system

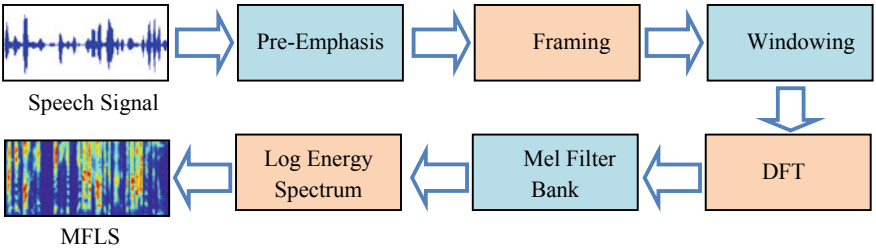

Fig. 2 Flow diagram of mel frequency log spectrogram (MFLS) process

$$\text{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{1}$$

Complete processing of MFCC is time-consuming, and the application of discrete cosine transform (DCT) provides a higher frequency resolution but a lower spatial relationship. Thus, we formulated part of MFCC that can be input to two-dimensional DCNN can maintain better frequency and spatial relationship [17].

Pre-emphasis. The pre-emphasis filter suppresses the random noise and amplifies the high-frequency components of the speech emotion signal. The equation for the pre-emphasis filter H(z) is given by Eq. 2.

$$H(z) = 1 - \beta.z \tag{2}$$

where β is a pre-emphasis coefficient that lies between 0 and 1.

Framing and Windowing. The speech emotion signal is non-stationary; therefore, to process stable speech components, it is split into frames of the 40 ms. To obtain a smooth changeover between frames, 50% overlapping of frames is used. Further, Hamming window is used to collect the closest frequency components together and avoid the leakage phenomenon. The Hamming window $W(n)$ for $\alpha = 0.46$ and N samples can be expressed by using Eq. 3.

$$W(n) = (1 - \alpha) - \alpha \cdot \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \tag{3}$$

Discrete Fourier Transform (DFT). DFT is used to transform the time-domain speech emotion signal into the frequency domain. The DFT $X(k)$ of the speech emotion signal $x(n)$ given is using Eq. 4.

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j2\pi nk/N}, \quad 0 \le n, k \le N-1 \tag{4}$$

The emotion power spectrum is by taking the square of the modulus of $X(k)$ as given in Eq. 5.

$$P(k) = \frac{1}{N}|X(k)|^2 \tag{5}$$

Mel Filter Bank. The mel spectrum can be obtained by passing the emotion power spectrum $P(k)$ through the mel-scale triangular filter bank. The product of $P(k)$ and $H_m(k)H_m(k)$ is computed at each frequency. Triangular filter bank frequency response $H_m(k)$ for $M = 32$ filters is computed using Eq. 6. We have considered $M = 32$, which can cover the frequency components between 133 and 3954 Hz.

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{6}$$

where $f(m)$ stands for the centre frequency of the mel frequency filter.

Mel Frequency Logarithmic Spectrum. The logarithmic energy spectrum $S(m)$ for each frame is computed using Eq. 7.

$$S(m) = \log_e\left(\sum_{k=0}^{N-1} P(k).H_m(k)\right), \quad 0 \le m \le M \tag{7}$$

where $P(k)$ represents the power spectrum, $H_m(k)$ is a triangular filter bank response, and M is a number of filters.

2.2 Deep Convolutional Neural Network

Two-dimensional CNN is popular for images processing ability which represents the internal correlation and saliency information of the local region of two- or three-dimensional image. It also helps to describe the spatial, temporal and frequency domain representation of two-dimensional data [18]. Mini batch gradient descent method is employed for the learning of the DCNN. In this architecture, each feature map is convolved with each kernel filter at every layer. Various layers of CNN are as follows:

Convolution Layer. In the convolution layer, the mel frequency spectrogram convolved with the convolution filter bank. Convolution layer describes the spatial local connectivity and correlation of the local region of the mel frequency spectrogram. In the convolution layer, two-dimensional mel frequency spectrogram $S(m)$ convolved with the convolution kernel $w(i, j)$ having a size $(p \times q)$ as given in Eq. 8. The weights of kernel $w(i, j)$ are initialized randomly.

$$C(i,\ j) = S(i,\ j) * w(i,\ j) = \sum_{m=0}^{p} \sum_{n=0}^{q} S(m, n).w(i - m,\ j - n) \tag{8}$$

Exponential Linear Unit (ELU). ELU removes the negative weights from the convolution layer output and normalizes the convolution layer output using Eq. 9.

$$E(i,\ j) = \begin{cases} C(i,j), & \text{if } C(i,j) > 0 \\ e^{C(i,j)}, & \text{if } C(i,j) \leq 0 \end{cases} \tag{9}$$

Maximum Pooling Layer (MP). The maximum pooling layer acts as a nonlinear function, and it only considers the salient information of non-overlapping local subregion. It increases the robustness of features against distortions and noise. Max pooling also helps to reduce the feature dimension. In this implementation, the maximum pooling window of 2×2 pixels is used with a stride of 2×2.

Fully Connected Layer (FC). The FC layer is similar to the multilayer perceptron network (MLP) that combines each neuron of a single layer to every neuron of other layers. The flattened output of the pooling layer is given to the FC layer as input.

Softmax Classifier. Softmax classifier is used for the multiclass emotion classification, which is the generalized framework of the logistic regression. Softmax function provides the probability of the predicted class (P_i), and the output class label (Y) is decided based on the maximum of P_i as given in Eq. 10–12.

$$P_i = \text{soft}\max(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_i}} \tag{10}$$

$$Y = \max_i(P_i) \tag{11}$$

$$z_i = \sum_{j}^{i} h_j.W_{ji} \tag{12}$$

where z_i is the output of an FC layer and input to the soft-max classifier, h_j is the activation function of the penultimate layer, and W_{ji} weight connecting penultimate and softmax layer.

Learning algorithm. For the learning of proposed DCNN, mini-batch gradient leaning method, which is a combination of stochastic gradient descent (SGD) and batch gradient descent (BGD) is used. It is robust and computationally cheaper [19]. In this, n samples are break up in smaller batches b. The error function for updating the weights (w) of DCNN is given in Eq. 13.

$$E_t[f(w)] = \frac{1}{b} \sum_{i=(t-1)b+1}^{tb} f(w, x_i) \tag{13}$$

where x_i is an ith sample of the training data, the weights are revised using mini-batch gradient update rule considering learning rate μ as given in Eq. 14.

$$w^{t+1} = w^t - \mu \nabla_w E[f(w^t)] \tag{14}$$

3 Experimental Results and Discussion

The proposed method is simulated using MATLAB software on a personal computer with a Core i5 CPU with 4 GB RAM on Windows environment. For the experimentation, Berlin Emo-DB speech emotion public database is used which consists of 535 utterances of 10 actors for seven emotions such as anger, boredom, fear, disgust, neutral, happiness and sadness [20]. The sampling rate used for data collecting is 16,000 Hz. We have considered the 5-s long speech samples to keep the uniformity in the mel frequency log spectrogram. If the sample length is less than 5 s, then it is padded to 5 s long using original signal. Otherwise, the samples are cropped to 5 s. The MFLS has the dimension of M $\times$ $F = 32 \times 249$ where M is number of triangular filter banks, and F is the number of frames for the 5-s signal for 40 ms duration and 50% frame-shift.

In the first CNN, input MFLS (32 $\times$ 249) convolved with the six kernels of 3 $\times$ 3 filter with the stride of one pixel and without zero-padding followed by ELU and max pooling. Each feature map of every layer is convolved with each filter kernel. The feature dimensions for various layers of DCNN are given in Table 1. The output layer has seven neurons that correspond to the output labels of seven emotions.

Table 2 shows the % accuracy for SER based on the speaker-dependent and speaker-independent approaches is considering the direct audio clip and MFLS as

Table 1 Details of feature map of various layers of proposed work

Layer	Sub-layer	Kernel size	Stride	Feature map
Input layer	Mel frequency log spectrogram	–	–	32 × 249
CNN layer 1	Convolution Layer 1	3 × 3 × 6	1	29 × 247 × 6
	ELU Layer 1	–	–	29 × 247 × 6
	Max Pooling Layer 1	2 × 2	2	14 × 123 × 6
CNN layer 2	Convolution Layer 2	3 × 3 × 6	1	12 × 121 × 36
	ELU Layer 2	–	–	12 × 121 × 36
	Max Pooling Layer 2	2 × 2	2	6 × 60 × 36
CNN layer 3	Convolution layer 3	3 × 3 × 6	1	4 × 58 × 216
	ELU layer 3	–	–	4 × 58 × 216
	Max pooling layer 3	2 × 2	2	2 × 29 × 216
FC Layer	–	–	–	12,528 × 1

Table 2 % accuracy for SER based on Emo-DB database

Emo-DB emotion	Speaker-dependent approach		Speaker-independent approach	
	Audio spectrogram as input	MFLS as input	Audio spectrogram as input	MFLS as input
Anger	94.49	100	96.06	100
Boredom	88.88	93.83	97.54	98.76
Disgust	84.78	97.82	78.19	91.31
Fear	89.85	94.2	94.2	97.1
Happy	89.43	91.55	73.83	93.66
Neutral	100	92.4	77.21	94.93
Sad	93.55	100	93.04	96.78
Average accuracy	91.56	95.68	87.15	96.07

Table 3 Comparison of the proposed method with the previous implementation based on % accuracy (Emo-DB)

Research work	Method	Speaker-dependent approach	Speaker-independent approach
Huang et al. [12]	CNN	88.30	85.20
Zhao et al. [15]	CNN-LSTM	95.33	95.89
Proposed work	MFLS + DCNN	95.68	96.07

the input. When simple speech spectrogram is provided to the system, it resulted in 91.56% and 87.15% accuracy for speaker-dependent and speaker-independent modes, respectively. While when MFLS is provided as an input to the system, it gives a better improvement in performance and results in 95.68 and 96.08% accuracy for speaker-dependent and speaker-independent modes, respectively.

When the proposed method performance is assimilated with other implementations on the Berlin Emo-DB database for speech emotion recognition based on % accuracy, it is noticed proposed method has given satisfactory results as shown in Table 3.

4 Conclusion and Future Scope

This paper has presented the SER system based on the mel frequency log spectrogram (MFLS) and three-layered sequential deep convolutional neural network (DCNN). MFLS extracts the salient information from the raw speech emotion signal, which further boosts the discriminative feature extraction ability of two-dimensional DCNN. The performance of the proposed system is estimated on the Emo-DB database considering speaker-dependent (SD) and speaker-independent

(SID) approaches. The proposed method has resulted in 95.68 and 96.07% accuracy for the speaker-dependent and speaker-independent approaches when MFLS is provided as the input to the proposed DCNN. Our future work consists of an investigation of a proposed method for noisy database and spontaneous emotion speech database.

References

1. Schuller BW (2018) Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun ACM 61(5):90–99
2. Khalil RA, Jones E, Babar MI et al (2019) Speech emotion recognition using deep learning techniques: a review. IEEE Access 7:17327–117345
3. Gunawan TS et al (2018) A review on emotion recognition algorithms using speech analysis. Indonesian J Elect Eng Inf (IJEEI) 6(1):12–20
4. Anagnostopoulos C, Iliou T, Giannoukos I (2012) Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif Intell Rev 43(2):155–177
5. Guidi A, Vanello N, Bertschy G, Gentili C, Landini L, Scilingo E (2015) Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients. Biomed Signal Process Control 17:29–37
6. Dellaert F, Polzin T, Waibel A (1996) Recognizing emotion in speech. In: Proceeding of fourth international conference on spoken language processing. ICSLP'96, vol 3. IEEE, pp 1970–1973
7. Zhou Y, Sun Y, Zhang J, Yan Y (2009) Speech emotion recognition using both spectral and prosodic features. In: Information engineering and computer science, pp 1–4
8. Haq S, Jackson P, Edge J (2008) Audio-visual feature selection and reduction for emotion classification. In: Proceedings of international conference on auditory-visual speech processing (AVSP'08), Tangalooma, Australia
9. Sonawane A, Inamdar M, Kishor B (2017) Sound based human emotion recognition using MFCC & multiple SVM. In: 2017 international conference on information, communication, instrumentation and control (ICICIC), pp 1–4. IEEE, Indore, India
10. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44(3):572–587
11. Jianwei N, Yanmin Q, Kai Y (2014) Acoustic emotion recognition using deep neural network. In: 9th international symposium on Chinese spoken language processing, pp 128–132. IEEE
12. Huang Z, Ming D, Qirong M, Yongzhao Z (2014) Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM international conference on multimedia, pp 801–804
13. Zheng Q, Yu J, Zou Y (2015) An experimental study of speech emotion recognition based on deep convolutional neural networks. In 2015 international conference on affective computing and intelligent interaction (ACII), pp 827–831. IEEE)
14. Badshah A, Jamil M, A., Nasir, R., Sung, W.: Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 international conference on platform technology and service (PlatCon), pp 1–5. IEEE (2017)
15. Zhao J, Xia M, Lijiang C (2019) Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed Signal Process Control 47:312–323
16. Bhangale K, Titare P, Pawar R, Bhavsar S (2018) Synthetic speech spoofing detection using MFCC and radial basis function SVM. IOSR J Eng 8(6):55–62
17. Zheng F, Guoliang Z, Zhanjiang S (2001) Comparison of different implementations of MFCC. J Comput Sci Technol 16(6):582–589
18. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. 1st edn, MIT Press, Cambridge, MA

19. Shende P, Dandwate Y (2020) Convolutional neural network based multimodal biometric human authentication using face, Palm Veins and Fingerprint. Int J Innov Technol Explor Eng (IJITEE) 9(3):771–777
20. Burkhardt F, Paescke A, Rolfes M, Sendlmeirer WF, Weiss B (2005) A database of German emotional speech. In: 9th European conference of speech commincation and Technology