

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Data augmentation using a 1D-CNN model with MFCC/MFMC features for speech emotion recognition

Thomas Mary Little Flower, Thirasama Jaya & Sreedharan Christopher Ezhil Singh

To cite this article: Thomas Mary Little Flower, Thirasama Jaya & Sreedharan Christopher Ezhil Singh (2024) Data augmentation using a 1D-CNN model with MFCC/MFMC features for speech emotion recognition, *Automatika*, 65:4, 1325-1338, DOI: [10.1080/00051144.2024.2371249](https://doi.org/10.1080/00051144.2024.2371249)

To link to this article: <https://doi.org/10.1080/00051144.2024.2371249>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 03 Jul 2024.



Submit your article to this journal [↗](#)



Article views: 1124



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



Data augmentation using a 1D-CNN model with MFCC/MFMC features for speech emotion recognition

Thomas Mary Little Flower ^a, Thirasama Jaya ^b and Sreedharan Christopher Ezhil Singh ^{b,c}

^aDepartment of ECE, St.Xavier's Catholic College of Engineering, Chunkankadai, India; ^bDepartment of ECE, Saveetha College of Engineering, Chennai, India; ^cDepartment of Mechanical Engineering, Vimal Jyothi Engineering College, Kannur, India

ABSTRACT

Speech emotion recognition (SER) is attractive in several domains, such as automated translation, call centres, intelligent healthcare, and human–computer interaction. Deep learning models for emotion identification need considerable labelled data, which is only sometimes available in the SER industry. A database needs enough speech samples, good features, and a better classifier to identify emotions efficiently. This study uses data augmentation to enhance the amount of input voice samples and address the data shortage issue. The database capacity increases by adding white noise to the speech signals by data augmentation. In this work, the Mel-frequency Cepstral Coefficient (MFCC) and Mel-frequency Magnitude Coefficient (MFMC) features, along with a one-dimensional convolutional neural network (1D-CNN), are used to classify speech emotions. The datasets utilized to estimate the model's enactment were AESDD, CAFE, EmoDB, IEMOCAP, and MESD. The data augmentation with the 1D-CNN (MFMC) model performed best, with an average accuracy of 99.2% for AESDD, 99.5% for CAFE, 97.5% for EmoDB, 92.4% for IEMOCAP and 96.9% for the MESD database. The proposed 1D-CNN (MFMC) with data augmentation outperforms the 1D-CNN (MFCC) without data augmentation in emotion recognition.

ARTICLE HISTORY

Received 19 December 2023
Accepted 18 June 2024

KEYWORDS

Neural networks; affective computing; emotion recognition; audio database; accuracy

1. Introduction

Speech is the most effective, widespread, and natural mode of human communication. A speech signal conveys information about the speaker's gender, age, language, dialect, and emotional state in addition to conveying a message [1]. One method that has emerged due to technological advancement to enhance and expedite human–computer contact is speaking to a machine. Because of this, researchers have looked into various strategies over the last few decades to improve spoken communication's efficacy using technologies like voice and speaker recognition [2]. One of the primary goals of speech emotion recognition is to generate a machine that can hear and react like a human and produce different types of emotions found in speech. Simulating the distinctive relationship between the information a microphone picks up and the accompanying emotion is the main challenge [3]. Speech can express emotions either consciously or unconsciously, thanks to the neurological system. Emotional speech recognition recognizes a person's emotional state from their voice. Due to its numerous uses in many fields, emotion identification is gaining prominence in the detection of frustration, disappointment, surprise, and amusement [4].

Spectral-based features like MFCC, log-frequency power coefficients (LFPC), log-mel spectrograms

(LMS), linear prediction coefficients (LPC), and MFCC [5–8] are among the most common feature types [9]. However, SER's poorer accuracy in detecting emotions remains a challenge. The difficulties arise because various people express emotions differently, standardized databases are available, and the same speech signals might convey multiple emotions based on the circumstances [10]. We investigated the feature significance for every classifier using feature importance methodologies [11]. MFMC and MFCC characteristics achieve speech emotion categorization in this study.

Deep learning (DL) has recently attracted much attention from the research community. The feature extraction procedure has been mechanized in DL [12]. As a result, hidden patterns can be successfully found even in the manually extracted features, enhancing the SER operation's performance. However, customized features have been very successful for SER. Researchers working on emotion recognition are paying much attention to deep learning algorithms [13]. Generally, DNNs perform better than regular neural networks (NNs). However, they frequently cause overfitting issues and need a lot of training samples to overcome them. Restricted Boltzmann Machines (RBMs), a DL approach developed initially, are used in the Deep Belief Network (DBN). DBN is faster than a

standard neural network because of the use of RBM. Later, convolutional neural networks (CNN) gained much traction because of their enhanced discriminative power over DBN. Convolution, pooling, tangent squashing, rectifier, and normalizing make up a standard CNN algorithm [14]. To build a progressive hierarchy of usable features, CNN uses feature extractions and a few convolutional stacks [15]. Convolutional layers and subsampling layers follow a hierarchical NN structure [16]. The 1D CNN model works well in time-series data and has shown tremendous potential for speech-emotion classification tasks. In this work, we have analyzed speaker-independent emotion recognition from speech using 1D-CNN (MFCC) and 1D-CNN (MFMC) models with and without data augmentation.

Many deep learning (DL) models have been searching to increase the accuracy of speech emotion identification, including deep CNNs, recurrent neural networks (RNN), and long short-term memory (LSTM) networks. Due to difficulties in collecting, speech emotion corpora are often tiny and deficient datasets. The performance of the DL models is constrained since they are prone to overfitting. This aims to identify the most beneficial kind of data augmentation and the quantity of data augmentations needed to solve the SER issue. This data augmentation technique permutes the original data with noise to generate fresh voice samples from a given dataset. This study examined 1D-CNN (MFCC) and 1D-CNN (MFMC) models with and without data augmentation for speaker-independent emotion identification from speech. The critical influence of this study is as given below:

- We thoroughly examined the literature on speech emotion recognition and discovered that speech emotions are hybrid, including spectral and spatial information. According to the literature, both of these traits provide crucial data for recognizing emotions. The present SER systems don't have DA with CNN architecture to learn high-level deep features and recognize the emotions present in the speech signal. Due to this restriction, we suggested two models for emotion recognition using 1D - CNN (MFCC) and 1D - CNN (MFMC) models with and without DA.
- This work is not only theoretical but also practical. We aim to categorize the speech emotion using a 1D CNN (MFCC) model and to detect the high-level hidden supra-segmental characteristics from those extracted segmental features during training. To ensure the robustness of our research, we have used multiple public speech emotion datasets, including AESDD, CAFE, EmoDB, IEMOCAP, and MESD. This diverse dataset allows us to achieve the highest detection accuracy, making our research more applicable in real-world scenarios.

- Data augmentation is a key aspect of our model training process. It not only enhances the number of training samples but also decreases overfitting, thereby improving the model's generalization ability. The model that was trained with 60% of data with DA shows a significant improvement in detection accuracy, validating the effectiveness of our approach.

This paper deliberates the details of the workflow. The sections are discussed as follows: The SER task's current literature review is presented in Section 2 to help readers understand the current trend, develop their perception, and identify areas for task improvement. Specify the summary of the architecture datasets, data augmentation methods, feature extraction procedure, suggested model, and model training in Section 3. A thorough justification of the experimental findings for the suggested speaker-independent individual model's 1D-CNN (MFCC) and 1D-CNN (MFMC) with data augmentation is provided in Section 4. Section 5 serves as our conclusion and discusses the issues facing SER research and potential future directions.

2. Literature review

Taiba Majid Wani et al. [2] have presented the distinguishing silent discriminants and pertinent speech emotion recognition characteristics. On spectrograms produced from the SAVEE dataset, CNN and the Deep stride CNN (DSCNN) were applied. The accuracy of DSCNN architecture exceeds CNN, which has a prediction accuracy of 87.8%. Dias Issa et al. [3] state that emotion identification from speech is one of the most challenging topics in data science. The architecture uses MFCC features to classify emotions. This framework achieves a recognition accuracy of 71.61% for RAVDESS, 95.71% for EMO-DB, and 64.3% for IEMOCAP audio categorization tasks.

Youddha Beer Singh et al. [4] proposed a technique for recognizing emotions in speech using 1D CNN and MFCC characteristics. This method is assessed using RAVDESS, a prominent public speech corpus. Additionally, average accuracy was reported to be higher (82.93%) compared to the current SER model with lower computing costs.

Jothimani et al. [9] preprocessed the speech signals before using the MFCC, ZCR, and RMS feature extraction techniques to dramatically increase emotion identification ability. A cutting-edge CNN is suggested for improved emotion categorization. Recognition accuracy was obtained for the databases RAVDESS (92.6%), CREMA (89.5%), SAVEE (84.9%), and TESS (99.5%).

Pan S-T and Wu H-J [12] provide a unique machine-learning model for speech emotion identification dubbed CLDNN, which integrates CNN, LSTM,

and DNN. The suggested model is experimentally evaluated using three databases: RAVDESS, EMO-DB, and IEMOCAP. The findings show that the LSTM model successfully represents the features recovered from the 1D CNN owing to the time-series nature of speech signals. Additionally, the data augmentation strategy used in this research improves the recognition accuracy and stability of the systems for various databases.

Vryzas et al. [17] have suggested a model for SER based on the CNN framework. The AESDD is the speech emotion dataset utilized to train and test the model, and DA methods are also used. The accuracy of the CNN architecture is 8.4% higher than that of the SVM baseline model.

Seknedj et al. [11] tested SER utilizing RAVDESS, EmoDB, and CaFE speech emotion datasets on three commonly utilized languages: English, German, and French. Four machine learning classifiers, such as SVM, Random Forest, Multi-Layer Perceptron, and Logistic Regression, were used. In terms of recognition rates and overall running performance, SVM was determined to be the best classifier, and its emotion recognition rate was 70.56%, 85.97%, and 70.61% for RADVESS, EmoDB, and café database, respectively.

To create samples for underrepresented emotions, Chatziagapi et al. [14] examined DA utilizing generative adversarial networks (GANs). Two datasets, IEMOCAP and FEEL-25k, were used to evaluate emotion recognition. The GAN-based technique shows a 10% relative performance gain in IEMOCAP and a 5% improvement in FEEL-25k when the minority classes are added. Multiple augmenting approaches were employed to supplement the training data and simplify the model demonstration, as discussed by Bautista et al. [15]. Mel-spectrograms were created from raw audio data and used as input for a CNN attention-based network. The test dataset achieves the maximum degree of accuracy, with 89.33% for a parallel CNN-Transformer network and 85.67% for a parallel CNN-BLSTM-Attention network on the RAVDESS dataset.

Atmaja et al. [16] experimented with the effects of DA techniques to enhance SER accuracy. The IEMOCAP and Twitter-based Japanese emotive speech datasets are used in the tests. The findings indicate speaker-independent data with two data augmentation with silence removal. The experiment shows the need to select the best DA approach for a given situation by highlighting the quantity of DA and the effectiveness of SER. To attend to emotional traits with various granularities, Xu et al. [18] executed a multiscale area attention in a DCNN. To increase the classifier's ability to generalize and execute DA via vocal length to address the issue of sparse data. On the IEMOCAP dataset, experiments produced results with a weighted accuracy (WA) of 79.34% and an unweighted accuracy of 77.54%. (UA). Using the IEMOCAP dataset, Etienne

et al. [19] described the CNN + LSTM architecture for SER. The methods of layer-wise optimizer tuning, batch normalization (BN) of recurrent layers, and data augmentation using vocal track length perturbation yield excellent outcomes of 64.5% for WA and 61.7% for UA on four emotions.

Jahangir R et al. [13] suggested a unique SER framework that uses data augmentation techniques to extract seven important feature sets from each speech. Using the EMO-DB database, the retrieved feature vector is fed into the 1D CNN to recognize emotions. The testing findings demonstrate that the SER framework outperformed existing SER frameworks with an accuracy of 96.7% for EMO-DB.

Md. Rayhan Ahmed et al. [20], prompted by the efficient feature extraction of CNN, LSTM, and GRU, offer an ensemble that uses the combined predictive performance of three distinct architectures. The ensemble model achieves 95.42% weighted average accuracy, which is state-of-the-art for EMO-DB datasets.

Liu et al. proposed a FaceNet model for SER in 2021. The spectrogram feature was extracted from the spoken signal and evaluated using the FaceNet model, which, owing to its clean signals, resulted in an emotion identification rate of 68.96% for the IEMOCAP database. The FaceNet model's spectrogram feature pretraining is practical according to the performance metrics.

Xu Yunfeng et al. [21] use a hierarchical-grained feature model (HGFM) to tackle poor emotion categorization. This model contrasts with a few baseline models, including the bidirectional contextual LSTM, the MDNN, the dialogue RNN, and the recurrent neural network. The findings demonstrate that HGFM performs well compared to baseline models.

Su et al. [22] used a graph attention approach on a gated recurrent unit network to discern emotions in voice inputs. For the IEMOCAP database, this strategy yielded a 63.8% emotion detection rate.

Saleem et al. [23] state that deep CNN combines 1D – and 2D-cNN models. Both temporal and spectral information were extracted using two parallel CNNs. The high-level features were concatenated and supplied as input to the Deep-CNN model to get the high-level features. This model used global feature learning (GRU) after CNN to examine contextual dependencies. This model obtained an accuracy of 94.2% for emotion identification in the EmoDB dataset and 81.1% in the IEMOCAP database by learning concatenated features and training faster than LSTM.

Zengzhao Chen et al. [24] investigated speech emotion identification using the attention mechanism (AMSNet), which combines frame-level manual characteristics with utterance-level deep features of varying weights. The two features' relative importance is controlled by the weight value, which assigns a higher weight to the more significant characteristics. The features increase the total recognition effect contribution

and improve the model's performance. To go above the limitations of traditional feature fusion, an autonomous training technique. The findings show that the accuracy of emotion recognition for IEMOCAP is 70.51%, whereas for EmoDB, it is 88.56%.

3. Overview of SER

a. Data Sets

The Mexican Emotional Speech Database (MESD) has ways to express anger (1), disgust (2), fear (3), happiness (4), neutrality (5), and sadness (6). The MESD has voices from both adult and young non-professional performers, including three females, two males, and six children. There were 864 separate utterances saved in 24-bit, 48kHz audio files [25–27].

The Canadian French Emotional (CAFE) Speech Dataset shows six primary emotions: anger (1), disgust (2), happiness (3), fear (5), surprise (6), sadness (7), and one neutral (4). Six male and six female actors each speak one line. The six fundamental emotions have two degrees of intensity: mild and robust. The 936 utterances comprise the dataset at a high-resolution 192 kHz sampling rate with 24 bits per sample [28].

Acted Emotional Speech Dynamic Database (AESDD) is an acted emotional speech database in the Greek language. This database has five different emotions: anger (1), disgust (2), fear (3), happiness (4), and sadness (5). About 500 emotional speech utterances were made from more than one recording. The recordings of the speech signal at a 44100 Hz sample rate with 24 bits (5 actors x 5 emotions x 20 utterances) [17, 29–31].

Emo-DB is a publicly available speech-emotion database in the German language. Five male and five female speakers contributed to the data recording. This database consists of 535 data samples and has seven emotions: anger (1), boredom (2), disgust (3), anxiety (4), happiness (5), sadness (6), and neutral (7) – the recordings of the speech signal with a 16 kHz sampling rate [32].

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is a standard multimodal and multispeaker (Male and Female) English language dataset. Five males and five females elicit emotions by reading from a script or improvising in a conversational setting, and the recordings of the speech signal with a 16 kHz sampling rate. There are nine emotions present in this dataset: happiness, anger, excitement, frustration, sadness, surprise, fear, neutral, and others. Each utterance can be divided into an improvised or scripted speech section and annotated into nine emotion labels. A total of five sessions are present in this database. This database consists of a total of 10,039 utterances [33].

In this work, a speech improvised dataset for emotion classification. This improvised dataset consists of

five sessions, each containing two speakers' speech samples. To evaluate the performance of the emotion recognition model, only four emotional categories, anger (1), sadness (2), happiness (3), and surprise (4) from improvised sessions 01, 02, and 03, are considered. This research used 759 improvised speech emotion samples for male and female speakers.

b. Data Augmentation

In this work for DA, Gaussian noise processing methods have been applied. The novelty of this work explains the experiment conducted with three different speech data augmentations (noise), resulting in 4 augmentation sets, as detailed in Table 1. Set 2 is obtained from Set 1 by permuting the speech signal with zero noise. Set 3 from Set 2 with a noise level of 0.01. Similarly, Set 4 from Set 2 with a noise level of 0.025 and Set 5 were obtained from Set 1 by permutation of the original speech signal with a noise level of 0.05 applied directly to the audio signal prior. In the first phase of DA [34–36], the original dataset (CAFE) was permuted from the dataset with zero noise to get 936 extra utterances. This process was applied, and the resulting dataset included 1872 utterances. In the second phase, the permuted first phase dataset was used with 0.01 noise to get extra 936 utterances; in the third phase, the permuted first phase dataset was used with 0.025 noise to get extra 936 utterances; and finally, the original dataset was permuted with 0.05 noise to get extra 936 utterances. After data augmentation, four thousand six hundred eighty speech samples were for the CAFÉ database.

Set 1 data samples = Original (fAn1)

Set 2 data samples (fDi2) = (fAn1) permutation with zero noise

Set 3 data samples (fFe3) = (fDi2) + 0.01 noise

Set 4 data samples (fHa4) = (fDi2) + 0.025 noise

Set 5 data samples (fNe5) = (fAn1) permutation + 0.05 noise

Total emotion samples = Set 1 + Set 2 + Set 3 + Set 4 + Set 5

The data augmentation [20,21] technique with several speech samples in Table 1 and augmentation processes applied to the five different speech emotion databases.

c. Feature Extraction

Feature extraction converts raw data into an executable form that shows the most critical information for that task. In voice recognition, characteristics are generally from the acoustic data. The MFCCs are influential in determining the spectral envelope of a signal, which is critical for characterizing various emotional states. Pitch, intensity, and energy are prosodic qualities

Table 1. Number of Speech Samples After Data Augmentation.

Database	Set 1	Set 2	Set 3	Set 4	Set 5	Total emotion data
AESDD	500	500	500	500	500	2500
CAFE	936	936	936	936	936	4680
EmoDB	535	535	535	535	535	2675
IEMOCAP	759	759	759	759	759	3795
MESD	864	864	864	864	864	4320

that describe the pattern of a spoken signal. Sometimes, prosodic cues cannot distinguish between emotions because they are too similar to be detected. The LPC, PLP, MFCC [37], and LFPC are a few algorithms used to denote the emotion of speech for the emotion recognition method. Feature extraction methods typically produce a multi-dimensional feature vector for all speech samples [19].

a. Mel-Frequency Cepstral Coefficients

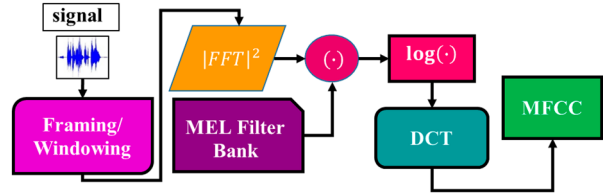
The speech sample is separated into short segments, as Figure 1 illustrates. Subsequently, speech analysis on the brief intervals known as frames, during which the speech signal remains stationary. The next step, windowing, tapers the signals close to the frame edges and provides spectral resolution for speech sounds. A hamming window to smooth the edges of the frames. The magnitude spectrum of each frame using the Fast Fourier Transform (FFT). Then, the Fourier spectrum passes through the mel-filter bank, and the mel-spectrum is obtained [16]. Mel-scale to circumvent the human auditory system's linear interpretation of pitch. Since people can detect subtle changes in speech at lower frequencies than at higher frequencies, it adjusts the frequency to roughly resemble what the human ear can perceive. When the Mel-scale is employed, the coefficients will only be around the region that humans perceive as the pitch, which may result in a more accurate demonstration of the signal from the viewpoint of the human emotion system when using the formula.

$$F_{mel} = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (1)$$

Where f is the natural frequency measured in Hz and F_{mel} is the observed frequency obtained on the mel scale, measured in mels, the mel spectrum is calculated by multiplying the power spectrum by triangular mel filters and in Equation (2).

$$P(m) = \sum_{k=0}^{N-1} (|X(k)|^2 H_m(k)); m = 0, 1, 2, \dots, M-1 \quad (2)$$

Where M is the number of triangular mel weighting filters $H_m(k)$ the k^{th} energy spectrum bin contributes the m^{th} output band and $|X(k)|$ is the magnitude spectrum. Next, the DCT is to translate the log-mel frequency coefficients into cepstral coefficients. MFCC is

**Figure 1.** Demonstrates the basic structure of MFCC.

the name of the conversion's output in Equation (3).

$$C[n] = \sum_{m=0}^{M-1} \log_{10}(P(m)) \cos \left(\frac{\pi n(m-0.5)}{M} \right); n = 0, 1, 2, \dots, L-1 \quad (3)$$

Where $C[n]$ are the cepstral coefficients and L is the number of MFCCs.

b. Mel Frequency Magnitude Coefficient

The MFMC feature extraction procedure is in Figure 2. Framing and windowing techniques are used first for each speech. The $X(k)$ spectra using a Fast Fourier transform on each windowed signal. To create the linear spectrum, which passes through several Mel-scale triangle filter banks, the relevant modulus and its square using the speech signal spectrum. The final step is to extract MFMC [38] characteristics by taking the logarithm on the sum of frequency components of the m^{th} band, which in Equation (4).

$$MFMC(m) = \log_{10} \left(\sum_{k=0}^{N-1} |X(k)| H_m(k) \right); m = 0, 1, \dots, M-1 \quad (4)$$

Where M is the number of triangular mel weighting filters $H_m(k)$ is the k^{th} spectrum bin contributing to the m^{th} output band and $|X(k)|$ is the magnitude spectrum.

c. Convolutional Neural Network

Figure 3 shows the importance of choosing the proper classifier after finding speech features. In the current investigation, emotions use deep learning methods such as CNN [24]. Since these networks use the mathematical function convolution, the word "convolutional" was coined. CNN is a DL system that uses emotion data as input, weights various characteristics of the emotion, and can distinguish between similar

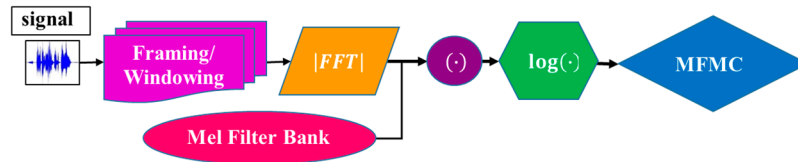


Figure 2. Demonstrates the basic structure of MFMC.

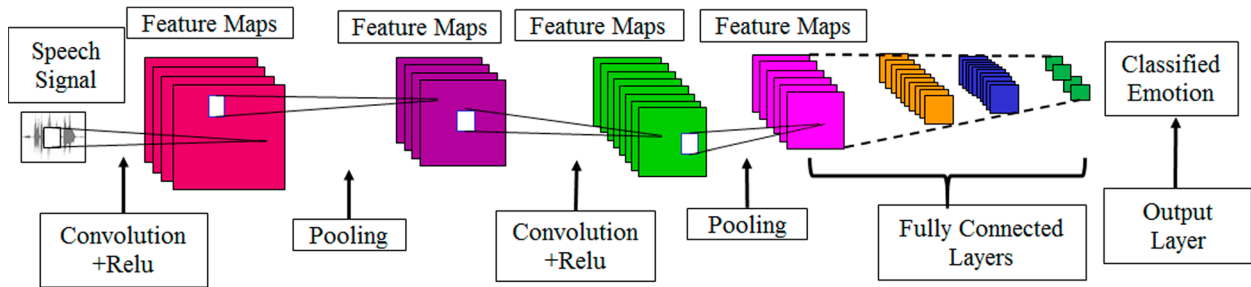


Figure 3. Demonstrates the CNN architecture.

emotions. The convolutional layer (CL), the pooling layer (PL), and the FC layer are typically the three components of a CNN. The input speech data in the input layer (IL) and the convolution layer calculate the output volume using the dot product between each filter and the emotion patch. This layer will apply an element-wise activation function to the CL output. The paramount persistence of the pooling layer, which is periodically added to CNN [22,23,39–41], is to decrease the volume, which speeds up computation and uses less memory. The FCL uses the data from the layer below to compute the class scores and produce a 1-D array with the same number of elements as classes.

d. Proposed Model of Speech Emotion Recognition

In the proposed model, the emotions classify to extract MFCC and MFMC features from the input speech corpus using the 1D CNN-MFCC and 1D CNN-MFMC models, as illustrated in Figure 4. Our suggested model uses 1D CNN and investigates emotion classification with and without data augmentation. The 1D CNN receives MFCC and MFMC features extracted from the speech samples. One hundred forty-four feature vectors as input data for our CNN's first layer. The first layer consists of stride 1, eight filters, and 15 kernel sizes. ReLU activates the output after the BN to a 1D Max Pooling (MP) layer with a window size of 1×2 . A second CL receives the output of the first IL with 16 filters that have the same kernel size as the first layer and the same stride as the first CL. ReLU activates the output after the BN is applied, and the output to a 1D MP layer with a window size 1×2 . The third CL receives the output from the second CL and contains 32 filters with the same size and stride. A 1D MP layer of the same size receives the output after ReLU activates it following the application of the BN. The fourth CL receives the output from the third CL and contains sixty-two

filters with identical sizes and strides. After applying the 0.3 dropout rate of the fourth CL and the 120 units of the FC layer, there is another 20% dropout rate with the seven units of the FC2 layer. ReLU activates the BN output to a one-dimensional MP layer of the same size. Depending on the number of anticipated classes, the output layer in the last stage uses a Softmax activation function.

e. Model Training

After acquiring the subset of features, the investigation calculates the emotion recognition accuracy by training the model with 60% of the data, 20% for testing, and the remaining 20% for validation. The augmented dataset to train models 1D-CNN (MFCC) and 1D-CNN (MFMC). The SER framework for DL throughout the entire procedure. The two models were trained separately for 100 epochs each. The best outcomes for the two models are when optimal weights to data augmentation.

4. Results and discussion

a. Confusion matrix for 1D CNN-MFCC without DA

The AESDD, CAFE, EmoDB, IEMOCAP, and MESD datasets for emotion recognition research. The proposed 1D CNN-MFCC model used these datasets and tested the effectiveness of the predictions. In the 1D CNN-MFCC model without augmentation, the overall accuracy of every emotion classification in the confusion matrix (CM). The CM shown in Figure 5 (a-e) classifies the emotion of class-wise accuracy at the diagonal, precision values at the bottom row, and recall values at every class's last column of the CM. From the result, anger emotion is misclassified mostly as happiness and disgust. The emotion recognition rate is higher

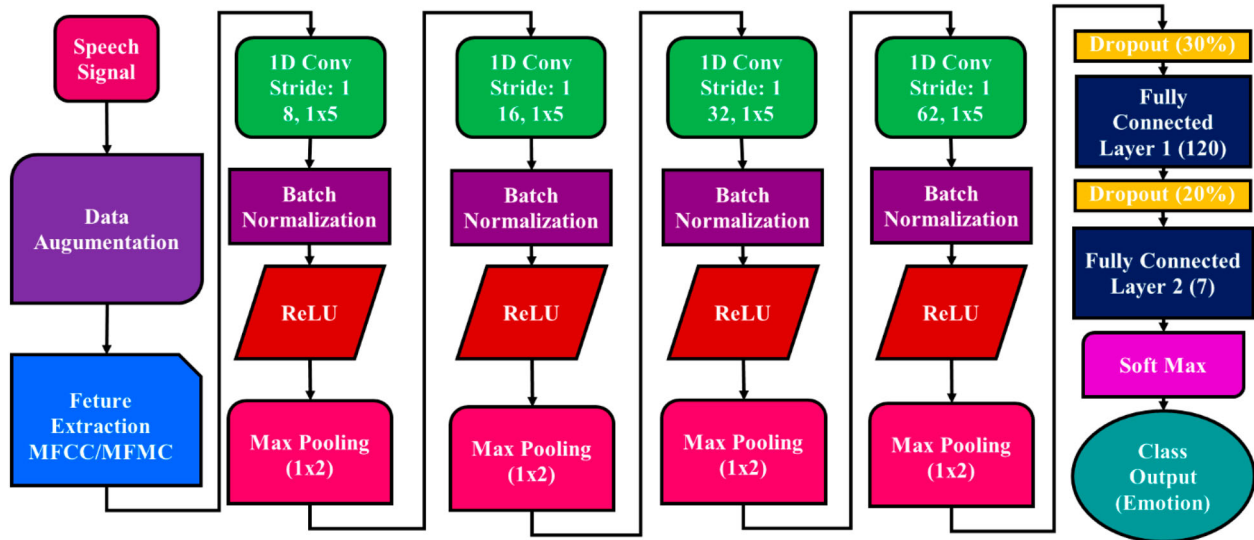


Figure 4. Demonstrates the proposed model with data augmentation.

for happy emotions. Classifying the associated speech signals is challenging for neutral emotions because of their less noticeable characteristics, contributing to their lower recognition rate. Learning rate is one of the most essential hyperparameters. The loss function varies more slowly as the learning rate decreases. However, if the learning set is too broad and far from the extreme value, the loss stops lowering and varies regularly at a certain point. Experiments in this study, and it was discovered that the learning rate should be 0.001. The figures in Figure 5 (a-e) demonstrate that the overall accuracy was 82% for the AESDD dataset, 88.8% for the CAFE dataset, 85% for the EmoDB dataset, 78.9% for the IEMOCAP dataset, and 85.5% for the MESD dataset as the emotion recognition for the 1D CNN-MFCC model without DA for SER.

b. Confusion matrix for 1D-CNN-MFCC with DA

The proposed 1D CNN-MFCC model with DA used the six datasets mentioned above and tested the efficiency of the predictions. In the 1D CNN-MFCC model with DA, the overall accuracy of emotion classification is shown in the CM. The emotion recognition rate is higher for boredom, surprise, and neutral emotions. The angry emotion is misclassified as happy due to the less observable features. The CM displayed in Figure 6 (a-e) classifies the emotion, and the overall accuracy was 95.2% for the AESDD dataset, 98.6% for the CAFE dataset, 96.1% for the EmoDB dataset, 91.2% for the IEMOCAP dataset, and 96.1% for the MESD dataset as the emotion recognition for the 1D CNN-MFCC model with DA for SER.

c. Confusion Matrix for 1D-CNN – MFMC without DA

The proposed 1D CNN-MFMC model without DA uses all six datasets and tests the effectiveness of the

predictions. In the 1D CNN-MFMC model without DA, the accuracy of every emotion classification is demonstrated in the confusion matrix (CM). The sad emotion is misclassified as a happy emotion due to the similarity of features. The CM shown in Figure 7 (a-e) classifies the emotion, and the overall accuracy was 95% for the AESDD test dataset, 93.6% for the CAFE test dataset, 86.9% for the EmoDB test dataset, 87.5% for the IEMOCAP test dataset and 94.2% for the MESD test dataset as the emotion recognition for the 1D CNN-MFMC model without DA for SER.

c. Confusion Matrix for 1D-CNN-MFMC with DA

The AESDD, CAFE, and MESD datasets were used in all six databases for emotion recognition. The proposed 1D-CNN-MFMC model with DA used these datasets and tested the efficiency of the predictions. In the 1D-CNN-MFCC model with DA, the overall accuracy of every emotion classification was demonstrated in the confusion matrix (CM). The result shows that most emotions are correctly predicted and have a reasonable emotion recognition rate using data augmentation and MFMC features. The CM shown in Figure 8 (a-e) classifies the emotion of class-wise accuracy, and the overall accuracy was 97.8% for the AESDD test dataset, 99.5% for the CAFE test dataset, 97.2% for the EmoDB test dataset, 92.4% for the IEMOCAP test dataset and 96.9% for the MESD test dataset as the emotion recognition for the 1D CNN-MFMC model with DA for SER.

e. Comparative Analysis of the Three Models with Other Studies

Table 2 compares the accuracy of speech emotion recognition of our proposed model with that of existing models. The accuracy of 1D-CNN (MFCC) and 1D-CNN (MFMC) models without and with data



Figure 5. (a) CM for 1D CNN-MFCC without DA for the database AESDD, (b) CAFÉ, and (c) EmoDB, (d) IEMOCAP, and (e) MESD for 30 coefficients.

Confusion Matrix : Test Dataset

1	96 19.2%	1 0.2%	2 0.4%	0 0.0%	0 0.0%	97.0% 3.0%
2	6 1.2%	94 18.8%	5 1.0%	2 0.4%	0 0.0%	87.9% 12.1%
3	1 0.2%	3 0.6%	83 16.6%	1 0.2%	0 0.0%	94.3% 5.7%
4	0 0.0%	0 0.0%	1 0.2%	96 19.2%	0 0.0%	99.0% 1.0%
5	0 0.0%	0 0.0%	1 0.2%	1 0.2%	107 21.4%	98.2% 1.8%
	93.2% 6.8%	95.9% 4.1%	90.2% 9.8%	96.0% 4.0%	100% 0.0%	95.2% 4.8%
	1	2	3	4	5	

(a)

Confusion Matrix : Test Dataset

1	132 14.1%	0 0.0%	2 0.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	98.5% 1.5%
2	0 0.0%	162 17.3%	0 0.0%	0 0.0%	0 0.0%	2 0.2%	0 0.0%	98.8% 1.2%
3	0 0.0%	0 0.0%	137 14.6%	0 0.0%	1 0.1%	0 0.0%	0 0.0%	99.3% 0.7%
4	0 0.0%	1 0.1%	0 0.0%	71 7.6%	0 0.0%	2 0.2%	0 0.0%	95.9% 4.1%
5	2 0.2%	0 0.0%	0 0.0%	0 0.0%	140 15.0%	0 0.0%	0 0.0%	98.6% 1.4%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	140 15.0%	0 0.0%	100% 0.0%
7	0 0.0%	3 0.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	141 15.1%	97.9% 2.1%
	98.5% 1.5%	97.6% 2.4%	98.6% 1.4%	100% 0.0%	99.3% 0.7%	97.2% 2.8%	100% 0.0%	98.6% 1.4%
	1	2	3	4	5	6	7	

(b)

Confusion Matrix : Test Dataset

1	113 21.1%	1 0.2%	0 0.0%	0 0.0%	4 0.7%	0 0.0%	0 0.0%	95.8% 4.2%
2	0 0.0%	94 17.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	46 8.6%	0 0.0%	0 0.0%	3 0.6%	0 0.0%	93.9% 6.1%
4	0 0.0%	1 0.2%	0 0.0%	61 11.4%	4 0.7%	0 0.0%	1 0.2%	91.0% 9.0%
5	0 0.0%	0 0.0%	0 0.0%	2 0.4%	60 11.2%	2 0.4%	0 0.0%	93.8% 6.2%
6	0 0.0%	0 0.0%	0 0.0%	2 0.4%	1 0.2%	69 12.9%	0 0.0%	95.8% 4.2%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	71 13.3%	100% 0.0%
	100% 0.0%	97.9% 2.1%	100% 0.0%	93.8% 6.2%	87.0% 13.0%	93.2% 6.8%	98.6% 1.4%	96.1% 3.9%
	1	2	3	4	5	6	7	

(c)

Confusion Matrix : Test Dataset

1	142 18.7%	2 0.3%	11 1.4%	6 0.8%	88.2% 11.8%
2	4 0.5%	403 53.1%	14 1.8%	6 0.8%	94.4% 5.6%
3	13 1.7%	1 0.1%	121 15.9%	4 0.5%	87.1% 12.9%
4	0 0.0%	6 0.8%	0 0.0%	26 3.4%	81.2% 18.8%
	89.3% 10.7%	97.8% 2.2%	82.9% 17.1%	61.9% 38.1%	91.2% 8.8%
	1	2	3	4	

(d)

Confusion Matrix : Test Dataset

1	138 16.0%	1 0.1%	3 0.3%	0 0.0%	0 0.0%	2 0.2%	95.8% 4.2%
2	0 0.0%	135 15.6%	2 0.2%	0 0.0%	0 0.0%	1 0.1%	97.8% 2.2%
3	2 0.2%	2 0.2%	145 16.8%	1 0.1%	1 0.1%	1 0.1%	95.4% 4.6%
4	2 0.2%	1 0.1%	0 0.0%	123 14.2%	3 0.3%	1 0.1%	94.6% 5.4%
5	0 0.0%	1 0.1%	0 0.0%	2 0.2%	148 17.1%	0 0.0%	98.0% 2.0%
6	3 0.3%	2 0.2%	0 0.0%	3 0.3%	0 0.0%	141 16.3%	94.6% 5.4%
	95.2% 4.8%	95.1% 4.9%	96.7% 3.3%	95.3% 4.7%	97.4% 2.6%	96.6% 3.4%	96.1% 3.9%
	1	2	3	4	5	6	

(e)

Figure 6. (a) CM for 1D CNN-MFCC with DA for the database AESDD (b) CAFÉ, (c) EmoDB, (d) IEMOCAP, and (e) MESD with 30 coefficients.

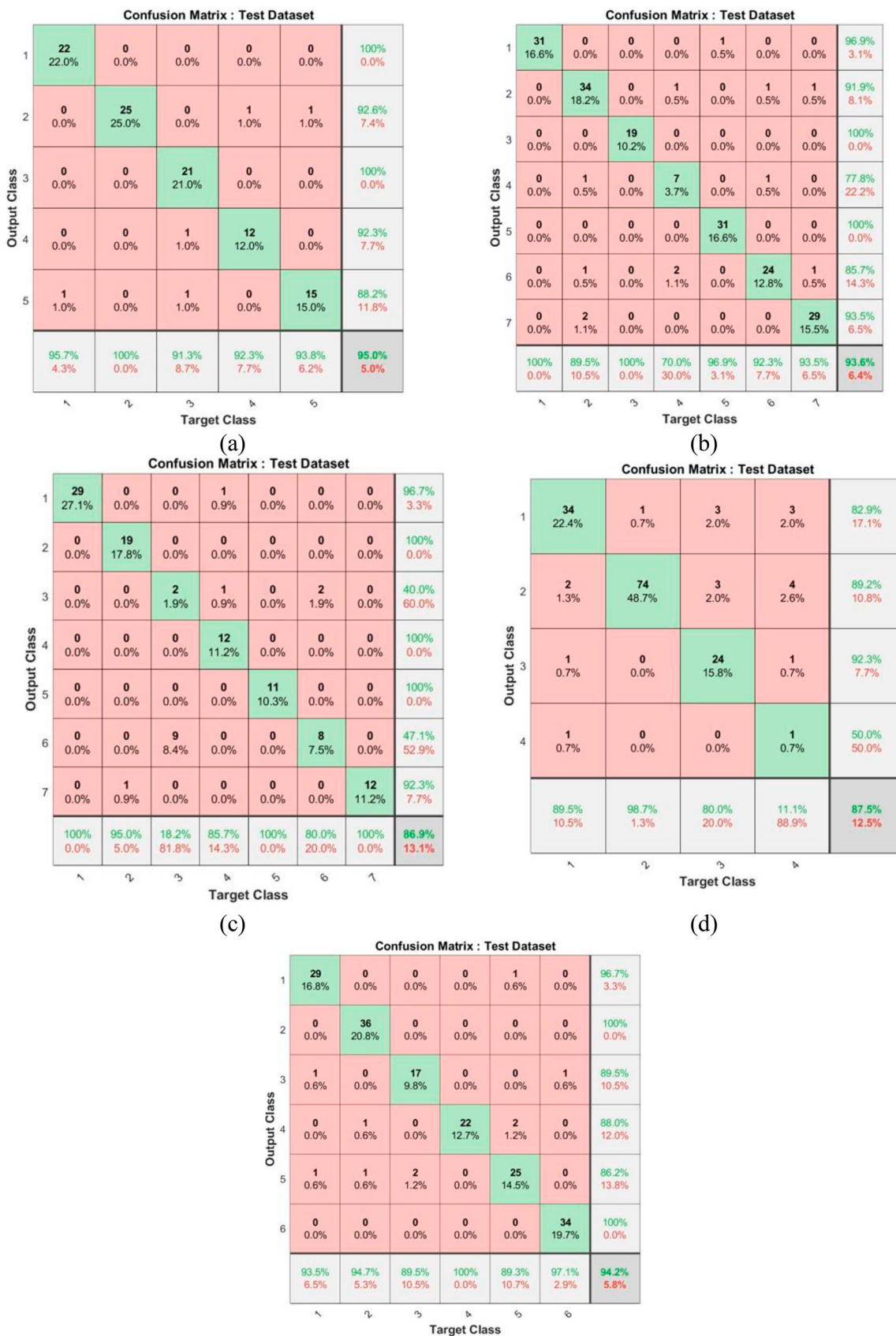


Figure 7. (a) CM for 1D CNN-MFMC without DA for the database AESDD (b) CAFÉ, (c) EmODB, (d) IEMOCAP, and (e) MESD with 30 coefficients.



Figure 8. (a) CM for 1D CNN-MFMC with DA for the database AESDD (b) CAFÉ, (c) EmoDB, (d) IEMOCAP, and (e) MESD with 30 coefficients.

Table 2. Compares the SER of our proposed models to that of the existing models.

Database	Reference	Accuracy (%)	Database	Reference	Accuracy (%)
Emo-DB	[11]	85.97	CAFÉ	[11]	70.61
	[42]	89.00		[42]	70.70
	[35]	73.46		[35]	47.01
	[36]	81.87		[36]	63.57
	[32]	91.25	IEMOCAP	proposed	99.50
	[22]	94.20		[32]	72.02
	[3]	86.10		[23]	81.10
	[12]	88.00		[3]	64.30
	[13]	96.70		[24]	70.51
	[20]	94.52		[22]	63.80
	[24]	88.56		[39]	68.96
	[7]	93.4		[21]	70.50
	[40]	77.8		[16]	76.39
	proposed	97.50		[18]	77.54
AESDD	[33]	87.10	[19]	61.70	
	[10]	68.00	proposed	92.40	
	[40]	70.00	MESD	[27]	88.90
	proposed	99.20	proposed	96.90	

Table 3. Comparative Analysis of the five models' accuracies.

Database	Coefficient	Accuracy (%)			
		1D-CNN – MFCC		1D-CNN – MFMC	
		without DA	with DA	without DA	with DA
AESDD	12	80	97.4	90	98
	24	89	97.8	96	99.2
	30	82	95.2	95	97.8
CAFÉ	12	85.6	94.3	92.5	96.7
	24	88.2	96.6	93	97.5
	30	88.8	98.6	93.6	99.5
EmoDB	12	85	93.5	90.7	96.64
	24	86	96.8	90.7	97.5
	30	85	96.1	86.9	97.2
IEMOCAP	12	77.6	88	84.2	92
	24	80.9	89.7	82.2	90.8
	30	78.9	91.2	87.5	92.4
MESD	12	82.7	91.4	90.2	91.67
	24	82.7	95.6	91.3	96.4
	30	85.5	96.1	94.2	96.9

augmentation is given in Table 3. The proposed 1D-CNN (MFCC) accuracy analysis with and without DA. The five datasets were used to assess the efficiency of each model, and 12, 24, and 30 coefficients were used for testing on all five data sets. The model data is trained 60%, 20% for testing, and 20% for validation. After training of 100 epochs, the 1D-CNN (MFCC) model is tested without and with data augmentation, producing 89% and 97.8% accuracy for the AESDD test dataset, 88.8% and 98.6% accuracy for the CAFÉ test dataset, 86% and 96.8% accuracy for the EmoDB test dataset, 80.9% and 91.2% accuracy for the IEMOCAP test dataset, and 85.5% and 96.1% accuracy for the MESD test dataset. The effectiveness of the 1D-CNN (MFMC) model is shown in Table 3 with and without the use of DA across all of the employed datasets. The five datasets were used to calculate the SER effectiveness of each model, and 12, 24, and 30 coefficients were used for testing on all five datasets. After training of 100 epochs, the 1D-CNN (MFMC) model is tested without and with data augmentation, producing 96% and 99.2% accuracy for the AESDD test dataset, 93.6% and 99.5% accuracy for the CAFÉ test dataset, 90.7%

and 97.5% accuracy for the EmoDB test dataset, 87.5% and 92.4% accuracy for the IEMOCAP test dataset, and 94.2% and 96.9% accuracy for the MESD test dataset. The 1D-CNN (MFMC) model shows enhanced accuracy for the data sets with DA compared to the 1D-CNN (MFMC) model.

5. Conclusions


In this research work, the two proposed models, 1D-CNN-MFCC and 1D-CNN-MFMC, are utilized to investigate speech emotion recognition accuracy for the five databases with data augmentation. The 1D-CNN (MFMC) model with DA performed better accuracy for recognizing emotion in all the five benchmark databases AESDD, CAFÉ, EmoDB, IEMOCAP, and MESD with an average model accuracy of emotion recognition for AESDD at 99.2%, CAFÉ at 99.5%, EmoDB at 97.5%, IEMOCAP at 92.4%, and MESD 96.9%. In future work, different feature extraction techniques will be used to classify emotions using augmentation techniques.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Thomas Mary Little Flower  <http://orcid.org/0000-0002-1163-3626>

Sreedharan Christopher Ezhil Singh  <http://orcid.org/0000-0002-3162-3847>

References

- [1] Uddin MZ, Nilsson EG. Emotion recognition using speech and neural structured learning to facilitate edge intelligence. *Eng Appl Artif Intel*. Sep 2020;94:103775. doi:10.1016/j.engappai.2020.103775
- [2] Wani TM, Gunawan TS, Ahmad Qadri SA, et al. Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks. 2020 6th International Conference on Wireless and Telematics (ICWT), Yogyakarta, Indonesia; Sep 2020. pp. 1–6. doi:10.1109/ICWT50448.2020.9243622.
- [3] Issa D, Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks. *Biomed Signal Proces*. May 2020;59:101894. doi:10.1016/j.bspc.2020.101894
- [4] Singh YB, Goel S. 1D CNN based approach for speech emotion recognition using MFCC features, Artificial Intelligence and Speech Technology. in Proc. 2nd International Conference on Artificial Intelligence and Speech Technology, (AIST2020), Delhi, India; Chapter 38, June 2020, pp. 1–8. doi:10.1201/9781003150664.
- [5] Ajibola Alim S, Khair Alang Rashid N. Some commonly used speech feature extraction algorithms. *From Nat Artif Intel Algorithms App*. Dec 2018. doi:10.5772/intechopen.80419
- [6] Ravi V, Wang J, Flint J, et al. Fraug: a frame rate based data augmentation method for depression detection from speech signals. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore; Apr 2022. pp. 6267–6271. doi:10.1109/ICASSP43922.2022.9746307.
- [7] Al-onazi BB, Nauman MA, Jahangir R, et al. Transfer-based multilingual speech emotion recognition using data augmentation and feature fusion. *Appl Sci*. Sep 2022;12(18):9188. doi:10.3390/app12189188
- [8] Badr Y, Mukherjee P, Thumati S. Speech emotion recognition using MFCC and hybrid neural networks. In Proceedings of the 13th International Joint Conference on Computational Intelligence (IJCCI 2021); Jan 2021, pp. 366–373. doi:10.5220/0010707400003063.
- [9] Jothimani S, Premalatha K. MFF-SAUG: multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network. *Chaos, Solitons Fractals*. Sep 2022;162:112512. doi:10.1016/j.chaos.2022.112512
- [10] Vryzas N, Vrysis L, Matsiola M, et al. Continuous speech emotion recognition with convolutional neural networks. *J Audio Eng Soc*. Feb 2020;68(1/2):14–24. doi:10.17743/jaes.2019.0043
- [11] Seknedy ME, Fawzi S. Speech emotion recognition system for human interaction applications. 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt; Dec 2021, pp. 361–368. doi:10.1109/ICICIS52592.2021.9694246.
- [12] Pan S-T, Wu H-J. Performance improvement of speech emotion recognition systems by combining 1D CNN and LSTM with data augmentation. *Electronics*. 2023; 12(11):2436. doi:10.3390/electronics12112436
- [13] Jahangir R, Teh YW, Mujtaba G, et al. Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion. *Mach Vision A*. Mar 2022;33(3). doi:10.1007/s00138-022-01294-x
- [14] Chatziagapi A, Paraskevopoulos G, Sgouropoulos D, et al. Data augmentation using GANs for speech emotion recognition. *Proc. Interspeech 2019*; Sep 2019, pp. 171–175. doi:10.21437/Interspeech.2019-2561.
- [15] Bautista JL, Lee YK, Shin HS. Speech emotion recognition based on parallel CNN-attention networks with multi-fold data augmentation. *Electronics*; 11(23):3935. doi:10.3390/electronics11233935
- [16] Atmaja BT, Sasou A. Effects of data augmentations on speech emotion recognition. *Sensors*. Aug 2022;22(16):5941. doi:10.3390/s22165941
- [17] Vryzas N, Liatsou A, Kotsakis R, et al. Augmenting drama: a speech emotion-controlled stage lighting framework. In Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences; Aug 2017, p. 8.
- [18] Xu M, Zhang F, Cui X, et al. Speech emotion recognition with multiscale area attention and data augmentation. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada; June 2021, pp. 6319–6323. doi:10.1109/ICASSP39728.2021.9414635.
- [19] Etienne C, Fidanza G, Petrovskii A, et al. CNN+LSTM architecture for speech emotion recognition with data augmentation. *Proc. Workshop on Speech, Music and Mind (SMM 2018)*, Sep 2018, pp. 21–25. doi:10.21437/SMM.2018-5.
- [20] Ahmed MR, Islam S, Muzahidul Islam AKM, et al. An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. *Expert Syst Appl*. May 2023;218:119633. doi:10.1016/j.eswa.2023.119633
- [21] Xu Y, Xu H, Zou J. HGFM : a hierarchical grained and feature model for acoustic emotion recognition. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain; 2020, pp. 6499–6503. doi:10.1109/ICASSP40776.2020.9053039.
- [22] Su B-H, Chang C-M, Lin Y-S, et al. Improving speech emotion recognition using graph attentive Bi-directional gated recurrent unit network. *INTER-SPEECH*. 2020. doi:10.21437/Interspeech.2020-1733
- [23] Pawar MD, Kokate RD. Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. *Mul Timed Tools Appl*. Feb 2021;80:15563–15587. doi:10.1007/s11042-020-10329-2
- [24] Chen Z, Li J, Liu H, et al. Learning multi-scale features for speech emotion recognition with connection attention mechanism. *Expert Syst Appl*. 2023;214:118943. doi:10.1016/j.eswa.2022.118943
- [25] Duville MM, Alonso-Valerdi LM, Ibarra-Zarate DI. Neuronal and behavioral affective perceptions of human and naturalness-reduced emotional prosodies. *Front Comput Neurosc*. Nov 2022;16:1022787. doi:10.3389/fncom.2022.1022787

- [26] Duville MM, Alonso-Valerdi LM, Ibarra-Zarate DI. The Mexican Emotional Speech Database (MESD): elaboration and assessment based on machine learning. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico; Dec 2021, pp. 1644–1647, doi:10.1109/EMBC46164.2021.9629934.
- [27] Duville MM, Alonso-Valerdi LM, Ibarra-Zarate DI. Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody. *Data*. Dec.2021;6(12):130. doi:10.3390/data6120130
- [28] Gournay P, Lahaie O, Lefebvre R. A Canadian French Emotional Speech Dataset. (1.1) [Data set]. ACM Multimedia Systems Conference (MMSys 2018) (MMSys'18), Amsterdam, Netherlands. Zenodo, June 2018. doi:10.5281/zenodo.1478765.
- [29] Vryzas N, Kotsakis R, Liatsou A, et al. Speech emotion recognition for performance interaction. *J Audio Eng Soc*. June 2018;66(6):457–467. doi:10.17743/jaes.2018.0036
- [30] Vryzas N, Vrysis L, Kotsakis R, et al. Speech emotion recognition adapted to multimodal semantic repositories. In 2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), IEEE; Sep 2018, pp. 31–35.
- [31] Vryzas N, Matsiola M, Kotsakis R, et al. Subjective evaluation of a speech emotion recognition interaction framework. In Proc. of the Audio Mostly 2018 on Sound in Immersion and Emotion, ACM; Sep 2018, p. 34.
- [32] Xu X, Li D, Zhou Y, et al. Multi-type features separating fusion learning for speech emotion recognition. *Appl Soft Comput*. 2022;130:109648. doi:10.1016/j.asoc.2022.109648
- [33] Pham NT, Nguyen SD, Thuy Nguyen VS, et al. Speech emotion recognition using overlapping sliding window and shapley additive explainable deep neural network. *J Inf Telecommun*. doi:10.1080/24751839.2023.2187278
- [34] Busso C, Bulut M, Lee CC, et al. IEMOCAP: interactive emotional dyadic motion capture database. *J Lang Resour Eval*. Dec 2008;42(4):335–359. doi:10.1007/s10579-008-9076-6
- [35] Ng AJB, Liu K-H. The investigation of different loss functions with capsule networks for speech emotion recognition. *Sci Program*. 2021;2021:Article ID 9916915. doi:10.1155/2021/9916915
- [36] López-Gil J-M, Garay-Vitoria N. Assessing the effectiveness of ensembles in speech emotion recognition: performance analysis under challenging scenarios. *Expert Syst Appl*. 2024;243:122905. doi:10.1016/j.eswa.2023.122905
- [37] Burkhardt F, Paeschke A. A database of German emotional speech. *Interspeech*. 2005;5:1517–1520. doi:10.21437/Interspeech.2005-446
- [38] Yalamanchili B, Samayamantula SK, Anne KR. Neural network-based blended ensemble learning for speech emotion recognition. *Multidim Syst Sign Process*. Aug 2022;33:1323–1348. doi:10.1007/s11045-022-00845-9
- [39] Liu S, Zhang M, Fang M, et al. Chih-Cheng Hung speech emotion recognition based on transfer learning from the FaceNet framework. *J Acoust Soc Am*. 2021;149:1338–1345. doi:10.1121/10.0003530
- [40] Pentari A, Kafentzis G, Tsiknakis M. Speech emotion recognition via graph-based representations. *Sci Rep*. 2024;14:4484. doi:10.1038/s41598-024-52989-2
- [41] Saleem N, Gao J, Irfan R, et al. DeepCNN: spectro-temporal feature representation for speech emotion recognition. *CAAI Trans Intel Technol*. 2023;8:401–417. <https://doi.org/10.1049/cit2.12233>.
- [42] Agarla M, Bianco S, Celona L, et al. Semi-supervised cross-lingual speech emotion recognition. *Expert Syst Appl*. 2024;237(Part A):121368. doi:10.1016/j.eswa.2023.121368