Improvement of Speech Emotion Recognition by Deep Convolutional Neural Network and Speech Features

Aniruddha Mohanty[1][0000−0001−9799−4088], Ravindranath C. Cherukuri[2], Alok Ranjan Prusty[3]

[1] CHRIST (Deemed to be University), Bangalore, Karnataks, India
aniruddha.mohantyh@res.christuniversity.in
[2] CHRIST (Deemed to be University), Bangalore, Karnataks, India
cherukuri.ravindranath@christuniversity.in
[3] DGT, RDSDE, NSTI(W), Kolkata, West Bengal, India
dralokrprusty@gmail.com

Abstract. Speech Emotion Recognition (SER) is a dynamic area of research which includes features extraction, classification and adaptation of speech emotion dataset. There are many applications where human emotions play a vital role for giving smart solutions. Some of these applications are vehicle communications, classification of satisfied and unsatisfied customers in call centres, in-car board system based on information on drivers' mental state, human-computer interaction system and others. In this contribution, an improved emotion recognition technique has been proposed with Deep Convolutional Neural Network (DCNN) by using both speech spectral and prosodic features to classify seven human emotions - anger, disgust, fear, happiness, neutral, sadness and surprise. The proposed idea is implemented on different datasets such as RAVDESS, SAVEE, TESS and CREMA-D with accuracy of 96.54%, 92.38%, 99.42% and 87.90% respectively and compared with other predefined machine learning and deep learning methods. To test the real time accuracy of the model it has been implemented on the combined datasets with accuracy of 90.27%. This research can be useful for development of smart applications in mobile devices, household robots and online learning management system.

Keywords: Emotion Recognition · Speech Features · Speech dataset · Data Augmentation · Deep Convolutional Neural Network.

1 Introduction

Emotions are the short-lived feeling which come from a notorious reason. It is observed as both mental and physiological state with the way of speaking, gestures, facial expressions etc. Speech can convey human emotions such as anger, disgust, fear, happiness, neutral, sadness and surprise. Same utterance can be observed based on different sounds and emotions. Speech emotion recognition

(SER) recognizes the emotional aspect of speech irrespective of the linguistic content which is more vital than studying the human emotions.

As per Ancilin and Milton [1] the speech features are segregated into prosodic and acoustic features. System features represent the response from the vocal tract. Das and Mahadeva [2] illustrated that source features represent periodicity, smoothed spectrum information and shape of the glottal signal.

Due to presence of many acoustic conditions like compressed speech, noisy speech, silent portion of the speech, telephonic conversation and so on, the existing SER approaches require more investigations with different acoustic conditional environments like intonation, stress, and rhythm.

Now-a-days due to availability of quality and variety of datasets, deep learning is the typical technique used in speech emotion recognition that helps a lot in the investigation. In this paper, an SER model have been proposed comprising of four different stages [3]. The first stage is preprocessing which includes pre-emphasis, framing, windowing, voice activity detection etc. The second stage is features extraction in which both spectral and prosodic features are considered. In third stage, feature selection and dimension reduction have been used. In final stage, deep learning concept that is DCNN have been applied in order to measure the performance of the model with various datasets like RAVDESS, SAVEE, TESS and CREMA-D. Further, the performance of the model is being evaluated in a realistic scenario by combining all the datasets.

The paper is organized in different sections. Section 2 is the Related Work which describes the existing emotion recognition techniques. Section 3 as the System Model that gives insight about the proposed idea and is like the black box representation. Section 4 describes the Proposed Method which deals with relevant features and features extractions, feature selection and the proposed procedure. Section 5 describes the Experiments and the Result analysis which shows the details about the used software, current work and their discussions. The summary and concluding remarks of the projected work is provided in Section 6.

2 Related Work

Abundant SER studies have been conducted over the years on emotion features, dimension reduction and classifications. Some other speech emotion recognition solutions are also proposed in the recent years.

Wang et al. [4], proposed a dual model where mel-frequency cepstral coefficients (MFCC) feature is processed using Long-Short Term Memory (LSTM) and Mel-Spectrograms is processed using Dual-Sequence LSTM (DS-LSTM) simultaneously. Issa et al. [5] proposed one-dimensional Convolutional Neural Network on the speech features like MFCC, chromagram, mel-scale spectrogram, Tonnetz representation, and spectral contrast features from speech files. Christy et al. [6] evaluated SER with the help of multiple machine learning algorithms like linear regression, decision tree, random forest, SVM and CNN. MFCC and modulation spectral (MS) are taken in the implementation. Again, Pawar and Kokate [7]

proposed one or more pairs of convolutions and max-pooling layers of CNN and implemented on speech features like Pitch and Energy, MFCC and Mel Energy Spectrum Dynamic Coefficients (MEDC). Jermsittiparsert et al. [8] developed deep learning based model called ResNet34 helped to recognize the speech words automatically to detect emotion with the help of MFCC, prosodic, LSP and LPC features. Bhangale and Mohanaprasad [9] proposed a three-layered sequential deep convolutional neural network (DCNN) on mel-frequency log spectrogram (MFLS) speech features and implemented with CNN and CNN-LSTM. Swain et al. [10] presented to extract 3-D log Mel-spectrograms features by first and second derivative of Mel-spectrograms, then a bi-directional-gated recurrent unit network with ensemble classifiers using Softmax and Support Vector Machine to get final classification.

Hence, prior studies differ in several ways like use of different datasets, various speech features and its combinations, several machine learning models of classification and environment of communication affecting the emotion recognition from speech. These gaps motivate to work on emotion identification from speech.

3 System Model

In the proposed approach, Deep Convolution Neural Network (DCNN) is used for the classification of the emotions after data augmentation and extracting features from speech. The model includes one-dimensional convolutional layers combined with max pooling, dropout and flatten layers. The output is activated by ReLU.

The system model embodies the various blocks of the proposed approach like black box representation shown in Fig. 1. The starting block is the speech sample, which is input to the model. Then data augmentation is done to achieve a set of realistic data for more visibility. The features like MFCC, Chroma STFT, Zero Cross Rate, Spectral Centroid and Mel Spectrogram are extracted. Finally, Convolution layer is applied with ReLu activation followed by max pooling and dropouts. Flatten is also used to get one dimensional data which creates a single long feature vector.

Hence, the model is fine-tuned by adding, removing or modifying the dropout rates in some of the layers based on the datasets to achieve the optimal results for this implementation as compared to the pre-defined models.

4 Proposed Method

The method presented in this implementation consists of both acoustic and periodic speech features, DCNN model with different datasets. To design the model, input data is required. So, the data is collected from various database.

Database: To verify the performance of the model, four different types of the databsets are used. Those are RAVDESS, TESS, SAVEE, CREMA-D shown in Table 1.

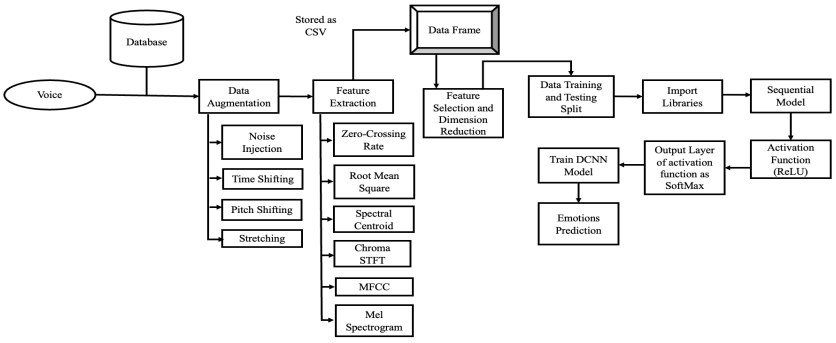

Fig. 1. Speech Emotion Recognition Approach

Table 1. Speech Corpora

Name	Language	Emotions	References
RAVDESS	English	calmness, happiness, sadness, anger, fear, surprise, disgust and neutral	[11]
SAVEE	English	surprise, happy, sad, angry, fear, disgust and neutral	[12]
TESS	English	happy, sad, angry, disgusted, neutral, pleasant surprise and fearful	[12]
CREMA-D	English	happy, sad, anger, disgust and neutral	[12]

4.1 Data Augmentation

Deep learning models heavily depend on training dataset. But the main drawback of the dataset is class imbalance and small size. To overcome these two drawbacks, data augmentation [8] is necessary to make the dataset more like real world audio inputs and enhance the recognition of emotions.

- **Noise injection** [12] is the technique which adds some arbitrary values into data.
- **Time shifting** [13] is the data augmentation technique in which audio is shifted to left or right with arbitrary seconds.
- **Pitch Shifting**[13] is based on time stretching and resampling technique.

4.2 Preprocessing and Feature Extraction

Several speech features are available to analyse the emotions from speech. Feature extraction plays an important role to implement any machine learning models. A leading trained model can be implemented by selecting the appropriate

features from speech samples [14]. So, several prosodic and spectral features representation of same sample is used as input to the deep learning model. Before extracting the prosodic and spectral speech features form speech samples, preprocessing needs to be completed.

Preprocessing [15] improves the intelligibility of the normal hearing system; helps to detect the highly powered speech segments of short durations; also detects the silent portion of the speech segments. Pre-emphasis filter is the first step as part of preprocessing which boosts the high frequency signals.

Feature Extraction Speech is a quasi-periodic signal of varying length which carries information and emotions. After preprocessing of speech signal, feature extraction can be performed which is the important aspect of emotion recognition.

Prosodic features are recognized by tones and the rhythms of the human voice. This is also called para-linguistic features which deals with the large units as phrases, words, syllables and sentence properties of speech elements. The used prosodic features are:

- **Zero-crossing rate (ZCR)** [14] is referred as rate of change of signal from positive to zero, from zero to negative and vice versa.
- **Root-Mean Square Value (RMS)** [14] measures the energy of a signal for each frame. The squares of each amplitude are added and divided by the number of amplitudes within frames.

Spectral features help to represent the characteristics of human vocal tract in frequency domain. In transferring the time domain signal into frequency domain by using Fourier Transform, spectral features [14] are obtained. Those are:

- **Spectral Centroid (SC)** [15] predicts the brightness of sound and referred as median of the spectrum.
- **Chroma STFT** [15] is used as short-term fourier transformation to calculate chroma features which helps to represent the arrangement of pitch and signal structure.
- **Mel Scale Spectrogram** [12] is a spectrogram where the frequencies are transformed to the mel scale. It is computed by following through Windowing, Fast Fourier Transform (FFT), generating a mel scale and generating spectrogram.
- **Mel Frequency Cepstral Coefficient (MFCC)** [12] is represented as short-term power spectrum of sound. It is derived with respect to the linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency.
- **Spectral Flux** [16] evaluates how quickly the power spectrum of a signal changes.

Feature Selection and Dimensionality Reduction In SER, feature selection helps to reach the best classification performance and accuracy from the feature set, also reduces training time and over-fitting [18].

Principal Component Analysis (PCA) [17] is an unsupervised learning dimensional reduction technique. In this method the feature sets are decomposed by covariance matrix to get principal components and weighted feature set. Then eigenvectors representative is selected based upon the weighted feature set.

4.3 Modelling

The Deep convolutional neural network (DCNN) [17] is a specialized type of multistage architecture neural network model designed to analyse emotions from speech shown in Fig. 2.

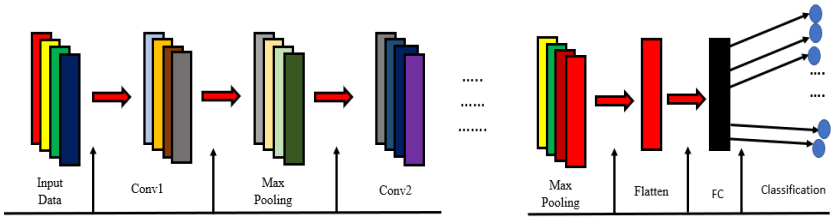

Fig. 2. Deep Convolutional Neural Network

Loss Function and Optimizer is one of the key aspects. The optimization algorithm continuously updates the direction of gradient descent by minimizing the loss function. Then each layer are updated by back propagation mechanism of neural network till the results are optimized. In the implementation, categorical cross entropy loss function [18].

After that, **Root Mean Squared Propagation (RMSProp)** [20] optimization algorithm is used to enhance the loss function. The RMSprop optimizer limits the fluctuations in the vertical direction. So, the learning rate of algorithm grew in horizontal direction.

4.4 Flowchart Emotion Classification

The flowchart shows the implementation of the proposed speech emotion recognition model which is represented in Fig. 3.

5 Experiment Set up

The proposed Speech Emotion Recognition model has been implemented by the help of Python language and its supported libraries along with various machine learning libraries. Python (Python 3.6.3rc1) along with librosa (librosa 0.8.0) is used for audio processing which provides various inbuild functions for implementation. Apart from these libraries, seaborn and matplotlib libraries are used

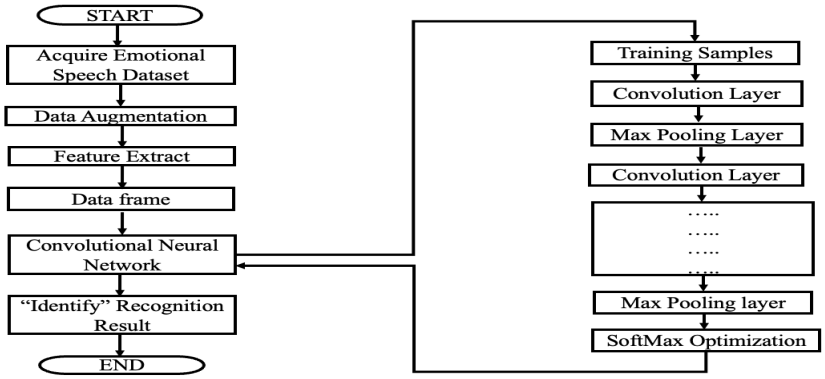

Fig. 3. Flowchat of the Implementation

to plot the graph which helps to analyse the data statistically. Machine learning libraries keras (keras-2.6.0), tensorflow (tensorflow-2.6.0) and Scikit-learn are also used in this implementation. The datasets are collected from the databases

Table 2. Layers and parameters of DCNN model

Layer (type)	Output Shape	Param#
RAVDESS		
dense (Dense)	(None, 14)	910
activation_8 (Activation)	(None, 14)	0
SAVEE		
dense (Dense)	(None, 7)	455
activation_8 (Activation)	(None, 7)	0
TESS		
dense (Dense)	(None, 7)	1351
activation_8 (Activation)	(None, 7)	0
CREMA-D		
dense (Dense)	(None, 12)	780
activation_8 (Activation)	(None, 12)	0
Combined datasets		
flatten (Flatten)	(None, 192)	0
dense (Dense)	(None, 12)	780
activation_8 (Activation)	(None, 12)	0

and loaded as input to the framework in the form of speech signal. Next, data augmentation techniques like Noise injection, Shifting and Pitch are used to update the dataset more similar to real-world speech input data. After the data augmentation, the dataset is stored as data frame in the CSV format. Each of the speech samples are preprocessed and speech features are extracted by the help of librosa library. The feature is selected by using Fast correlated based

filter where threshold value is used as 0.01. To get more accurate result, the dimension of the dataset have been reduced by the help of PCA. Then the deep learning concept, Deep Convolution Neural Network is applied for classification. In this implementation, the proposed model have been analysed with the used four datasets individually and again combining all the datasets. Each dataset is divided 75% as training data and 25% as testing data. The fine tune Layers, Output shape of each layer and parameters of DCNN are described in Table 2 for each dataset.

5.1 Result and Analysis

In order to classify emotions [12] and measure the performance of the designed model, four datasets (RAVDESS, SAVEE, TESS, CREMA-D) are used. The confusion matrix, precision, recall, and F1 score give better intuition of prediction results and accuracy is also compared. In the classification task, the obtained outputs are either positive or negative. There are four category of predictions such as False positive (fp), False negative (fn), True positive (tp) and True negative (tn) [12]. False positive refers as the model predicted samples as positive but actually negative. False negative corresponds as model predicts the samples as negative but actually positive. True positive is the sample where both prediction and actual are positive. The prediction and actual samples are negative for True negative.The results are analysed with the help of Precision, Recall, F1 Score, Support, Macro Average and Weighted Average values.
Confusion Matrix is a performance metric to measure a classification model where output is binary or multiclass having the table of different combinations. The Confusion matrix analysis of RADVESS, SAVEE, TESS, CREMA-D and the combined datasets are given in Fig. 4, Fig. 5 and Fig. 6.

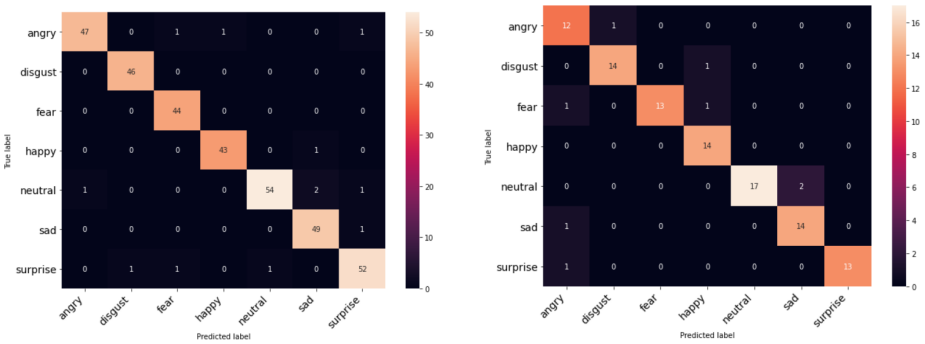

Fig. 4. Confusion matrix of RADVESS and SAVEE datasets

The accuracy measure for RADVESS, SAVEE. TESS. CREMA-D datasets are showing 0.97 (As 0.9666 is rounded), 0.93 (As 0.9253 is rounded), 0.99 (As

Table 3. Result analysis of the model with various datasets

Measures	Precision	Recall	f1 Score	Support#
Angry	0.98 (R)	0.94 (R)	0.96 (R)	50 (R)
	0.85 (S)	0.85 (S)	0.85 (S)	13 (S)
	1.00 (T)	1.00 (T)	1.00 (T)	98 (T)
	0.89 (C)	0.94 (C)	0.91 (C)	283 (C)
	0.89 (P)	0.88 (P)	0.89 (P)	90 (P)
Disgust	0.98 (R)	1.00 (R)	0.99 (R)	46 (R)
	0.88 (S)	1.00 (S)	0.99 (S)	15 (S)
	0.99 (T)	0.99 (T)	0.99 (T)	94 (T)
	0.88 (C)	0.88 (C)	0.88 (C)	329 (C)
	0.88 (P)	0.91 (P)	0.89 (P)	468 (P)
Fear	0.96 (R)	1.00 (R	0.98 (R)	44 (R)
	1.00 (S)	0.93 (S)	0.97 (S)	15 (S)
	1.00 (T)	0.99 (T)	0.99 (T)	94 (T)
	0.87 (C)	0.86 (C)	0.87 (C)	303 (C)
	0.86 (P)	0.93 (P)	0.89 (P)	470 (P)
Happy	0.98 (R)	0.98 (R)	0.98 (R)	44 (R)
	1.00 (S)	1.00 (S)	1.00 (S)	14 (S)
	1.00 (T)	1.00 (T)	1.00 (T)	106 (T)
	0.84 (C)	0.84 (C)	0.84 (C)	23 (C)
	.91 (P)	0.86 (P)	0.89 (P)	503 (P)
Neutral	0.98 (R)	0.93 (R)	0.96 (R)	58 (R)
	1.00 (S)	1.00 (S)	1.00 (S)	19 (S)
	0.98 (T)	0.99 (T)	0.99 (T)	106 (T)
	0.88 (C)	0.88 (C)	0.88 (C)	290 (C)
	0.89 (P)	0.97 (P)	0.93 (P)	446 (P)
Sad	0.94 (R)	0.98 (R)	0.96 (R)	50 (R)
	1.00 (S)	0.93 (S)	0.97 (S)	15 (S)
	1.00 (T)	0.99 (T)	1.00 (T)	104 (T)
	0.91 (C)	0.88 (C)	0.89 (C)	333 (C)
	0.91 (P)	0.85 (P)	0.88 (P)	499 (P)
Surprise	0.95 (R)	0.95 (R)	0.95 (R)	55 (R)
	1.00 (S)	1.00 (S)	1.00 (S)	14 (S)
	0.99 (T)	1.00 (T)	0.99 (T)	98 (T)
	0.97 (P)	0.92 (P)	0.95 (P)	466 (P)
Accuracy	-	-	0.97 (R)	347 (R)
	-	-	0.96 (S)	105 (S)
	-	-	0.99 (T)	700 (T)
	-	-	0.88 (C)	1861 (C)
	-	-	0.90 (P)	3342 (P)
M. Avg W. Avg	0.97 (R)	0.97 (R)	0.97 (R)	347 (R)
	0.96 (S)	0.96 (S)	0.96 (S)	105 (S)
	0.99 (T)	0.99 (T	0.99 (T)	700 (T)
	0.88 (C)	0.88 (C)	0.88 (C)	861 (C)
	0.90 (P)	0.90 (P)	0.90 (P)	3342 (P)

Note: R← RAVDESS, S ← SAVEE, T ← TESS, C ← CREMA-D, P ← PROPOSED.

0.9942 is rounded), 0.88 (As 0. 8790 is rounded) respectively. Hence the proposed model classifying perform as 97%, 93%, 99% and 88% of classification accuracy. Similarly, for combined dataset, the proposed model performance is showing 0.90 (As 0.9026 is rounded) and classification accuracy is 90%. TESS dataset which has only female speech samples shows extremely better performance in

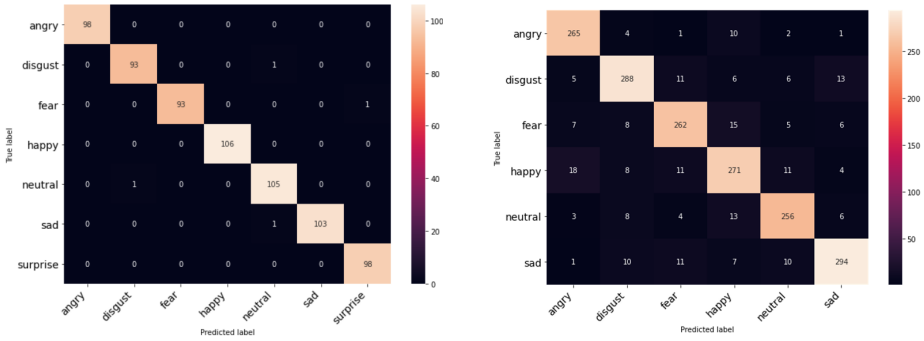

Fig. 5. Confusion matrix of TESS and CREMA-D datasets

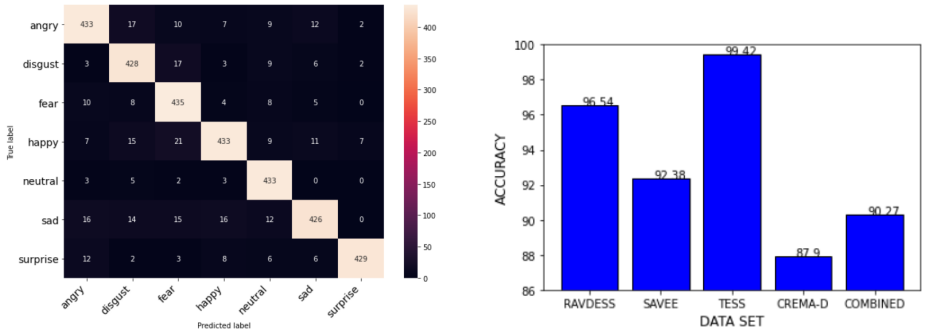

Fig. 6. Confusion matrix of Combined Dataset and Accuracy Values

this model. The detail analysis of the performance measures are discussed in Table 3.

5.2 Comparison Analysis

This implementation have been compared with previously implemented model which is using MFCC, Mel-scaled spectrogram, chromagram, spectral contrast feature and Tonnetz representation as speech features with CNN implementation. Even also compared with another model where ZCR, MFCC, Chroma STFT, RMS, Mel-scaled spectrogram are used as speech features with CNN modelling shows better performance illustrated in Table 4.

Various speech features are playing import ant role to improve the accuracy of the Emotion Recognition System, On considering other speech features such as Spectral centroid, Chroma STFT, Spectral flux as the combination of both prosodic and spectral features, the model performance for TESS is relatively better than other datasets. However, in this proposed model, it is observed that TESS dataset shows slightly less accuracy as compared to the existing model.

Table 4. Comparison with previous models

Dataset	Existing Model	Proposed Model
RAVDESS	71.76 [5], 88.72 [12]	96.54
SAVEE	65.03 [19], 86.80 [12]	92.38
TESS	55.71 [19], 99.52 [12]	99.42
CREMA-D	71.69 [12]	87.90
Combined Datasets	NA	90.27

6 Conclusion

It has been observed that deep learning technique like DCNN is a feasibility way of predicting human emotions from speech. The proposed model is composed of four phases - data augmentation, feature extraction, feature selection & dimensionality reduction and finally classification. DCNN is applied for classification of emotions on four different datasets like RAVDESS, SAVEE, TESS, CREMA-D. The performance of the model is verified by combining all the datasets into a single dataset and the proposed model is validated based on accuracy, precision, recall, micro average and weighted average values for different datasets. It has been observed that the proposed model outperforms the existing models with pre-defined machine learning and deep learning techniques. In future, the emotion classification accuracy of proposed model can be improved by considering different speech features like Spectral Spread, Spectral Entropy, Spectral Roll off, Chroma Vector, Chroma Deviation with other classification techniques.

References

1. Ancilin, J. & Milton, A.: Improved speech emotion recognition with Mel frequency magnitude coefficient. Applied Acoustics, 179, 1080469 (2021).
2. Das, Rohan Kumar & Mahadeva Prasanna, SR.: Exploring different attributes of source information for speaker verification with limited test data. The Journal of the Acoustical Society of America, 140(1), 184–190 (2016).
3. Daneshfar, Fatemeh, Kabudian, Seyed Jahanshah, & Neekabadi, Abbas: Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier. Applied Acoustics, 166, 107360 (2020).
4. Wang, Jianyou, Xue, Michael, Culhane, Ryan, Diao, Enmao, Ding, Jie, & Tarokh, Vahid: Speech emotion recognition with dual-sequence LSTM architecture. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6474–6478. IEEE, Barcelona (2020).
5. Issa, Dias, Demirci, M. Fatih, & Yazici, Adnan: Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894 (2020).
6. Christy, A., Vaithyasubramanian, S., Jesudoss, A., & Praveena, MD.: Multimodal speech emotion recognition and classification using convolutional neural network techniques. International Journal of Speech Technology, 23(2), 381–388 (2020).

7. Pawar, Manju D. & Kokate, Rajendra D.: Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. Multimedia Tools and Applications, 80(10), 15563–15587 (2021).
8. Jermsittiparsert, Kittisak, Abdurrahman, Abdurrahman, Siriattakul, Parinya, Sundeeva, Ludmila A., Hashim, Wahidah, Rahim, Robbi, & Maseleno, Andino: Pattern recognition and features selection for speech emotion recognition model using deep learning. International Journal of Speech Technology, 23(4), 799–806 (2022).
9. Bhangale, Kishor & Mohanaprasad, K.: Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network. Futuristic Communication and Network Technologies, 241–250 (2022).
10. Swain, Monorama, Maji, Bubai, Kabisatpathy, P., & Routray, Aurobinda: A DCRNN-based ensemble classifier for speech emotion recognition in Odia language. Complex & Intelligent Systems, 1–3 (2022).
11. Xu, Mingke, Zhang, Fan, & Zhang, Wei: Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset. IEEE Access, 9, 74539–74549 (2021).
12. Dolka, Harshit, VM, Arul Xavier, & Juliet, Sujitha: Speech Emotion Recognition Using ANN on MFCC Features. In: 2021 3rd International Conference on Signal Processing and Communication (ICPSC), pp. 431–435. IEEE, Coimbatore (2021).
13. Pham, Nhat Truong, Dang, Duc Ngoc Minh & Nguyen, Sy Dzung: Hybrid Data Augmentation and Deep Attention-based Dilated Convolutional-Recurrent Neural Networks for Speech Emotion Recognition. arXiv preprint arXiv:2109.09026, 309. 145–156 (2021).
14. Akçay, Mehmet Berkehan, & Oğuz, Kaya: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56–76 (2020).
15. Rajesh, Sangeetha & Nalini, NJ (2020), Musical instrument emotion recognition using deep recurrent neural network. Procedia Computer Science, 167, 16–25 (2020).
16. Hao, Yiya, Küçük, Abdullah, Ganguly, Anshuman, & Panahi, Issa MS: Spectral Flux-Based Convolutional Neural Network Architecture for Speech Source Localization and Its Real-Time Implementation. IEEE Access, 8, 197047–197058 (2020).
17. Zheng, WQ, Yu, JS & Zou, YX: 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, Xi'an (2015).
18. Gao, Mengna, Dong, Jing, Zhou, Dongsheng, Zhang, Qiang, & Yang, Deyun: End-to-end speech emotion recognition based on one-dimensional convolutional neural network. In: Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence, pp. 78–82. ACM Press, Kunming (2019). .
19. Mekruksavanich, Sakorn, Jitpattanakul, Anuchit, & Hnoohom, Narit: Negative emotion recognition using deep learning for Thai language. In: 2020 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON), pp. 71–74. IEEE, Pattaya (2020).
20. Tieleman, Tijmen and Hinton, Geoffrey, & others: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2), 26–31 (2012).