ISSN: 2584-0495 Vol. 1, Issue 1, pp.9-17

International Journal of Microsystems and IOT

ISSN: (Online) Journal homepage: https://www.ijmit.org

Speech Emotion Recognition for multiclass classification using Hybrid CNN-LSTM

Neha Prerna Tigga, Shruti Garg

Cite as: Neha Prerna Tigga, & Shruti Garg. (2023). Speech Emotion Recognition for multiclass classification using Hybrid CNN-LSTM. International Journal of Microsystems and IoT, 1(1), 9-17. https://doi.org/10.5281/zenodo.8158288

Published online: 26 June 2023.

DOI: https://doi.org/10.5281/zenodo.8158288

International Journal of Microsystems and IoT
Vol. 1, Issue 1, pp.9-17; DOI: https://doi.org/10.5281/zenodo.8158288

Speech Emotion Recognition for multiclass classification using Hybrid CNN-LSTM

Neha Prerna Tigga[1]*, Shruti Garg[1]

[1]Birla Institute of Technology, Mesra, Ranchi, India

ABSTRACT

Emotions are biological states of the human nervous system recorded in different signal forms that may be audio signals, electroencephalogram signals, etc. In this paper, cross-corpus emotion recognition is carried out on voice data. Also, a hybrid CNN–LSTM (Convolution Neural Network–Long Short-Term Memory) model was proposed for recognizing gender-biased emotions. Three established corpora were considered, namely, SAVEE, RAVDESS and TESS. Three new corpora have been constructed by combining the above-mentioned corpora for cross-corpus implementation, referred to as mix corpus. Corpora formed were gender-specific (i.e., male and female) and gender independent. Seven different emotions (i.e., happiness, sadness, anger, fear, neutral, disgust and surprise) have been identified within all the corpora. Data augmentation has been applied to reduce over-fitting and increase the robustness of deep neural networks by adding noise and pitch features to the signals. Also, the Mel-Frequency Cepstral Coefficient (MFCC) method was used for extracting feature before applying the hybrid network to each database. The experiment results show that the female corpus gives better accuracy than the male corpus.

KEYWORDS

Speech emotion recognition; CNN; LSTM; MFCC; Cross-corpus; Deep Learning

1. INTRODUCTION

Speech emotion recognition (SER) is a type of human–computer interaction and has undergone increased scrutiny recently [1]. Emotions are distinct and powerful mental actions, which may be perceived by various communicative behaviors [2]. Speech, motion, visual communication, brain signals, etc. signify the body's emotional state. Speech may be a fast, effective, and basic pathway of human communication [3]. The SER systems are extensively utilized in call centers, criminal investigation, neurology, identification psychiatric problem, etc. [4]. For different emotions, the audio signal varies in pitch, frequency, intensity, speaking tempo, and sound quality. [5]. Detecting the emotional intention of the mind is a typical task and can be used as a benchmark for any emotion identification method [6]. Among the various methods utilized for labelling these emotions, a discrete emotional technique is treated as one of the critical methodologies [7]. It incorporates different classes for emotions, i.e., anger, boredom, disgust, surprise, fear, joy, happiness, neutrality and sadness.

Table 1 below shows the work conducted in previous years using traditional and deep learning approaches, respectively.

1.1 Contribution

This work presents a hybrid CNN–LSTM technique to recognise emotions in audio signals. To pursue the research, the following steps has been taken:

1. Three databases, SAVEE, RAVDESS and TESS, were used for experimentation purposes.
2. Three new corpora have been created by combining audio files from the above-mentioned corpora. Two corpora were formed based on gender, referred to as gender-specific male and female corpora. Another corpus referred to as mixed corpus was created for a generalized cross-corpus implementation. The motivation behind adding the signals of three databases is to add variety in data.
3. Seven emotions were identified in all the databases, namely, happy, sad, anger, fear, neutral, disgust and surprise, by using the proposed hybrid CNN–LSTM model.

Our research is novel in two ways. First, in the conventional SER, training and testing is done on a single dataset, whereas this approach combines three datasets for emotion recognition. Second, the classification accuracy is also tested on a combination of different datasets based on gender. Third, multiclass classification is done in this research.

This paper is split into the following sections: Section 1 provides a background and introduction to SER utilizing deep learning techniques. The approach used in this study is described in Section 2. The experimentation is shown in Section 3. The fourth section contains the findings as well as a discussion. With supporting notes, Section 5 brings this effort to a conclusion.

Table. 1 List of studies on ser using deep learning methods

Paper / Year	Database	Emotions	Feature extraction	Deep learning approaches	Recognition accuracy
2023 [8]	IEMOCAP	Sad, neutral, happy, anger	MFCCs, mel spectrogram	CNN-BLSTM	Up to 717.70%
2023 [9]	TESS, EMO-DB, RAVDESS, SAVEE, CREMA-D	Disgust, surprise, happy, sad, anger, fear, calm, boredom	Time domain, frequency domain and spectral	1D-CNN-LSTM-GRU	Up to 99.46%%
2023 [10]	EmoDB, eNTERFACE05	Disgust, surprise, happy, sad, anger, fear	Chaogram images	Deep CNN	Up to 96.04%
2022 [11]	EmoDB, eNTERFACE, CASIA	Anger, bored, disgust, fear, happy, sad, neutral	Hierarchical alignment layer in feature extractor block	domain invariant feature learning	Up to 88.49%
2021 [12]	Berlin, RAVDESS, SAVEE, EMOVO, eNTERFACE, Urdu	Anger, surprise, disgust, fear, joy, sad neutral	Spectral features, Mel Frequency Magnitude Coefficient	Multiclass SVM	Up to 95.25%.
2021 [13]	RAVDESS	sad, happy, angry, fear, surprised, neutral, disgust	MFCC, GFCC, combined features	Deep C-RNN	More than 80%.
2020 [14]	AMIGOS	Neutral, disgust, happy, surprise, anger, fear, and sad	Spectrogram	Network of BLSTM, LSTM-RNN and DNN	Valence accuracy of 73.9%

**(GFCC,* Gammatone *Frequency Cepstral Coefficients; C-RNN, convolution-recurrent neural networks; AMIGOS, affect, personality traits and mood on Individuals and GrOupS; RML, Ryerson Multimedia Research Lab; BAUM-1; Bahçeşehir University Multimodal Affective Database; IEMOCAP, Interactive Emotional Dyadic Motion Capture)*

2. METHODOLOGY

The common SER system consist of following steps shown in figure 1.

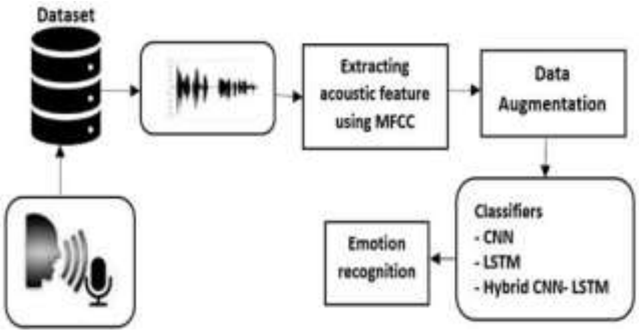

Fig. 1 Components of SER

2.1 Feature extraction

Speech signal requires processing to expel noise before the extraction and identification of significant features from speech. The point of the extracting feature is to outline a speech signal with predefined segments of speech. This is all because of the inconvenience of handling acoustic features. Feature extraction alters the speech segments to a short representation that is more differential and dependable than the original speech segment.

The present work implements MFCC for feature extraction, which is by far the most popular and established approach in the field of speech recognition [15]. The sound produced by speech is refined by the structure of the vocal tract, which decides what sound is generated. The vocal tract shows itself engulfed in a speech signal in a brief power spectrum for sound. The unit for measuring a pitch or frequency of a signal is Mel. Formula for conversion from a given frequency (f) to Mel is:

$$Mel\ (f) = 2595 * log_{10}\ \left(1 + \frac{f}{700}\right) \qquad (1)$$

Formula for conversion from frequency back to Mel is,

$$Mel^{-1}(m) = 700\ \left(exp\left(\frac{m}{1125}\right) - 1\right) \qquad (2)$$

MFCC can give better resolution frequency in poor frequency areas and is composed of seven steps, as given in Figure 2.

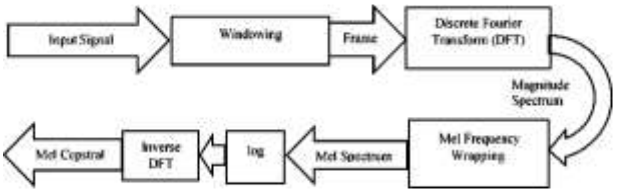

Fig. 2 Steps of MFCC

2.2 Hybrid CNN-LSTM

The CNN is formed by sharing weighted sums, convolution filters and pooling layers. Furthermore, it is a composition of several layers stacked together. The architecture of CNN varies according to three components (i.e., convolution filters and the pooling fully connected layers).

The features extracted from MFCC are fed into the CONV 1D convolution layer where the learning takes place. To induce nonlinearity in the structure, the ReLU activation function is applied in all the convolution layers. Non-linearity makes it simple for the data to adjust with varied types of data and distinguish between the outputs. The pooling layer is used to simplify the feature vector by preserving only the most significant features. The SoftMax activation function is employed in the fully connected layer. The final layer, reduces the previous layer's feature vector to a vector with a predetermined number of classes (in this case, seven emotions). The convolution theorem for one-dimensional CNN can be given as:

$$\begin{aligned} convoluted\ (x) &= f \otimes g \\ &= \int_{-\alpha}^{\alpha} f(x-u)\ g(u)\ du \\ &= F^{-1}(\sqrt{2\pi}\ F\ [f]\ F[g]) \qquad (3) \end{aligned}$$

Where f and g are the general continuous function, ⊗ defines the convolutional operator, F is the Fourier transformation and F^{-1} is the inverse F.

LSTM is a recurrent neural network (RNN) that learns from longstanding contextual reliability. Unlike different neural networks with feedforward connection, LSTM has a feedback connection. It is capable of handling sequence prediction problems such as speech recognition. Input, output, forget, and memory cells are the four components of the LSTM memory cell.

The equation listed below represents the process of updating the memory cells for each timeslot t. It all starts with the computation of the input gate value, denoted by ipt, and the candidate value for different stages of memory cell, given by Cmt, at a given timeslot t:

$$ip_t = \sigma\ (w_i\ x_t + u_i\ h_t\ +\ b_i) \qquad (4)$$

$$Cm_t = \sigma\ (w_c\ x_t\ +\ u_c\ h_{t-1}\ +\ b_c) \qquad (5)$$

Following the computation of the activation function for the new memory cell that forgets at timeslot t:

$$f_t = \sigma\ (w_f x_t + u_f h_{t-1} + b_f) \qquad (6)$$

From the obtained new value from the activation function of input gate ipt, forget gate ft and the candidate state value Cmt, the new state of the memory cell, Ct, can be calculated as:

$$C_t = ip_t * Cm + f_t * C_{t-1} \qquad (7)$$

With this obtained memory state cell, output gate values can be calculated as follows:

$$Op_t = \sigma\ (w_o x_t + u_o h_{t-1} + v_o C_t + b_o) \qquad (8)$$

$$h_t = Op_t * tan\,h\ (C_t) \qquad (9)$$

Where xt is the input to memory at timeslot t; wi, wf, wc, wo, ui, up, uc, uo and vo are the associated weights; and bi, bf, bc, bo are the corresponding biases. A CNN–LSTM model was proposed here by varying layers to improve the system's recognition accuracy. The combination of CNN and LSTM is a fruitful task, as they overcome each other's problems and make a strong classifier. It combines CNN's advantage of great prediction and LSTM's advantage of sequencing. LSTM adds memory to the CNN model, which increases the prediction capabilities. The proposed CNN–LSTM architecture is given below in Figure 3.

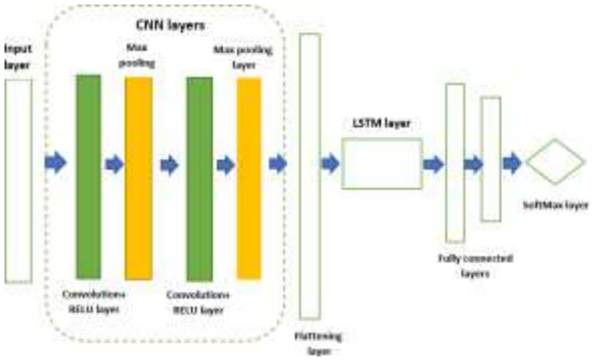

Fig. 3 Hybrid CNN-LSTM architecture

3. EXPERIMENTATION

This section reviews the databases considered, the applied feature extraction technique and the classification method used to improve SER. The deep learning classifiers used for speech recognition in this work are CNN, LSTM and hybrid CNN–LSTM. The feature extraction method implemented is MFCC. Figure 4. shows the workflow of our research.

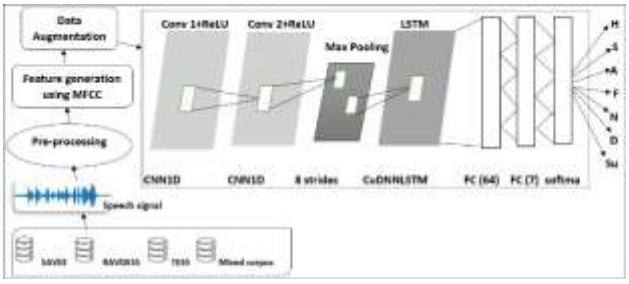

Fig. 4 Workflow of our research

3.1 Dataset description

3.1.1 SAVEE

The proposed technique was implemented in three benchmark corpora. First, this study utilised the SAVEE database, an English dataset developed in UK for automated speech recognition, collected at the University of Surrey [16]. The dataset was created using four male actors expressing seven different emotions (happiness, sadness, anger, fear, neutral, disgust and surprise) that have been recorded in audio–visual format. In total, 480 different statements have been recorded. This work only considers the speech part of the database.

3.1.2 RAVDESS

The second database is RAVDESS [17], a US English database, which is a collection of recordings of audio–video samples of speech and song from 24 actors (12 males and 12 females) expressing eight different emotions (happiness, sadness, anger, calm, fear, neutral, disgust and surprise). The emotion 'calm' has been removed from the database to compare across databases with only seven emotions. Only audio speech samples are considered for this experiment (in this case, 1440 audio files).

3.1.3 TESS

The third database utilised is TESS [18], collected by the University of Toronto. It consists of speech data spoken by two female actors (one young and the other old) articulating seven emotions (happiness, sadness, anger, fear, neutral, disgust and surprise). There are 2800 audio statements recorded in total.

3.1.4 Gender specific dataset

The fourth and fifth databases were created based on gender (i.e., male and female, respectively). The male corpus was created by combining SAVEE and RAVDESS (male). Total speech instances for the male corpus amounted to 3600 after data augmentation to match the number of instances from the female corpus. For the female corpus, the RAVDESS (female) and TESS databases were combined. Total speech instances totaled 3520 instances.

3.1.5 Mixed dataset

The sixth database was created by mixing all the above databases irrespective of gender and without data augmentation (i.e., SAVEE + RAVDESS + TESS) for the seven emotions. This brought the total audio samples to 4720 for the mixed corpus.

3.2 Data preprocessing

Pre-processing is a crucial step; the duration of the speech signals must be uniform, and the speech duration of the speech signals has been retained for all signals by trimming the edges. The audio length is trimmed to 2.5 seconds for every speech signal. Normalisation has been performed for missing value replacement.

3.3 Data augmentation

To overcome the problem of over-fitting and improve the robustness of a classifier, data augmentation was used. Data augmentation is most used on audio waveforms, and it involves making slight changes to the original waveform in order to generate a newer waveform. This helps increase the size of the database to serve as a good input for deep learning classifiers. Injecting noise, changing pitch and time and speed changes are some of the ways to augment a database. This work only incorporates two augmentation methods to the original databases – one by introducing noise and the other by changing the pitch. Then, the original database and both sets of augmented databases were combined and fed to the classifier for learning. Augmentation was only applied on the SAVEE and RAVDESS databases, as they had fewer audio instances. The final database size became 1440*216 for SAVEE and 1560*216 for RAVDESS.

3.4 Feature Extraction and classification

The 216 features are extracted after data pre-processing and augmentation using the MFCC feature extraction method. LibROSA is a package provided by Python for evaluating music/audio files. mfcc(), a feature extraction function of the LibROSA package, is used to extract features from the audio signals. This work extracted 216 features; thus, the dimension of SAVEE was 480*216, RAVDESS was 1440*216, TESS was 2800*216, female corpus was 3520*216, male corpus was 3600*216 and the mixed corpus was 4720*216.

These features are subsequently fed into two convolutional layers that use the ReLU activation function to process them. To minimize the number of features, maximum pooling is used. Here the stride is set to 8, which means the filter will move eight units at a time. After down-sampling, the features are fed into an LSTM layer. The output from LSTM then passed from to two fully connected layers flattens the feature map into vectors. Finally, to categorize the seven emotions as output, a SoftMax activation function is used.

A gender-biased and gender-neutral emotion recognition is done in present work. Each corpus was subdivided into an

80:20 ratio, with 80% of the data being used for training and 20% being utilized for testing. Experiments were conducted on Google Colab using the Python programming language to accelerate processing. The Keras open-source Python library was used to implement deep learning architectures, and a CONV 1D Keras layer was applied to a CNN layer. A CuDNNLSTM layer was applied to use LSTM. CuDNNLSTM provides faster implementation of LSTM using GPU processing.

All the layers were assigned ReLU activation functions except for the output layer, which was assigned SoftMax activation. The Adam optimizer was used throughout the experiment. The batch size was 16, and early stopping was applied to avoid over-training and reduce over-fitting. The model stopped once there was no improvement in accuracy. 'Dropout = 0.1' was used as a regularization method while training the network.

4. RESULT AND DISCUSSION

Three categories of experiments were conducted on all six corpora by varying the number of layers: firstly, by applying CNN classifier; secondly, applying LSTM and thirdly, implementing hybrid CNN–LSTM. Tables 2 and 3 show the best recognition accuracy obtained across all databases.

Table. 2 Recognition accuracies for single corpus (%)

	Model	Layers	A	D	F	H	N	S	Su	AVG
SAVEE	CNN	5	88.89	91.89	88.24	76.47	98.61	94.59	78.95	89.58
	LSTM	5	82.93	59.38	82.86	66.67	84.72	84.85	59.52	75.69
	CNN - LSTM	5	97.56	84.38	85.71	99.99	94.44	96.97	80.95	**91.66**
RAVDESS	CNN	5	83.33	81.58	86.00	69.81	92.86	77.08	93.62	82.69
	LSTM	5	78.26	76.19	65.38	62.75	82.14	62.75	66.67	69.55
	CNN - LSTM	5	95.65	90.48	84.62	70.59	85.71	88.24	88.10	**85.89**
TESS	CNN	9	84.25	89.83	84.97	94.33	93.46	93.43	85.31	89.33
	LSTM	5	93.15	86.44	83.66	91.49	97.39	96.35	79.72	89.61
	CNN - LSTM	7	91.50	91.49	94.41	97.08	97.39	95.89	89.83	**93.80**

**(A, anger; D, disgust; F, fear; H, happy; N, neutral; S, sad; Su, surprise; AVG, average percentage)*

Table. 3 Recognition accuracies for mixed corpus (%)

Database	Model	Layers	A	D	F	H	N	S	Su	AVG
SAVEE + RAVDESS (male)	CNN	7	45.88	53.85	44.90	28.70	81.76	48.39	57.78	53.19
	LSTM	7	35.29	40.66	39.80	34.78	74.32	37.63	35.56	44.86
	CNN - LSTM	7	49.41	51.65	38.78	45.22	85.81	55.91	41.11	**54.86**
RAVDESS + TESS (female)	CNN	9	79.25	74.45	61.78	75.55	92.37	68.47	80.33	**76.16**
	LSTM	9	80.19	73.57	56.44	77.73	86.86	73.42	63.18	73.01
	CNN - LSTM	9	83.96	70.93	57.33	67.69	88.14	66.67	84.10	74.21
Mixed Corpus	CNN	9	68.11	69.94	55.79	65.52	72.07	53.87	72.25	**65.62**
	LSTM	9	63.79	62.20	60.06	57.47	77.83	58.29	57.80	63.21
	CNN - LSTM	9	55.48	66.67	57.01	63.51	76.76	61.33	63.29	64.25

**(A, anger; D, disgust; F, fear; H, happy; N, neutral; S, sad; Su, surprise; AVG, average percentage; Mixed database, (SAVEE+ RAVDESS+TESS))*

It shows the highest accuracy with the corresponding layer and the recognition accuracy obtained for each emotion. Neutral was the highest predicted emotion out of all seven emotions. Hybrid CNN–LSTM performed the best in gender-specific individual corpora.

CNN was the second-best performing model. LSTM was the most average performing model in all the experiments. All the classification models yielded recognition accuracy between 42% and 93.8% across the database. The best performing male-based individual database was SAVEE with the highest accuracy of 91.66%, followed by RAVDESS with 85.89% accuracy via the hybrid CNN–LSTM classifier. When considering the female-based individual database, TESS performed best with 93.80% accuracy via hybrid CNN–LSTM, followed by RAVDESS with 49.44% via hybrid CNN–LSTM.

Extensive experiments were conducted on gender-specific and mixed corpora. Again, hybrid CNN–LSTM was the best performing model for mixed gender-specific corpora. For male corpora (SAVEE+RAVDESS), the highest accuracy of 54.86% and for female corpora (RAVDESS+TESS), the highest accuracy of 76.16% were both obtained by hybrid CNN–LSTM. In another experiment, all three databases were mixed together irrespective of gender and the three classifiers were applied. Here, CNN was little a better. The best performing classifier was CNN, producing an accuracy of 65.62%, followed by CNN–LSTM with 64.25% accuracy. It was noted that hybrid CNN–LSTM performed the best on individual gender-specific corpora and mixed gender-specific corpora. CNN and hybrid CNN–LSTM performed very similarly in the case of mixed corpora, with a difference of approximately 1%. Numerous experiments have been conducted by various researchers to improve recognition accuracy for speech. Some of the contemporary research has been compared to this study in Table 4 for each corpus.

The field of speech emotion recognition has witnessed significant advancements and discussions in recent years. Researchers have explored various techniques and methodologies to accurately recognize and classify emotions from speech signals.

One important aspect of the discussion revolves around the choice of feature extraction methods. Traditional approaches used handcrafted features such as MFCCs, pitch, and energy. However, with the advent of deep learning, researchers have explored the use of deep neural networks to automatically learn discriminative features directly from raw speech signals. CNNs have shown promising results in capturing local patterns and spectral information, while recurrent neural networks (RNNs) such as LSTM and GRU have proven effective in modeling temporal dependencies [19].

Another key area of discussion is the influence of dataset selection and size on model performance. Large, diverse datasets enable better generalization and robustness of the models. Additionally, the bias and imbalance in existing datasets can impact the performance of emotion recognition systems, leading to skewed results.

Using mixed corpora in SER allows for a more comprehensive and diverse representation of emotions. Traditional approaches often rely on single-domain or single-language corpora, which may limit the generalizability of the trained models. In contrast, mixed corpora encompass a wide range of emotions expressed by speakers from various demographics, cultures, and languages, providing a more realistic and robust representation of real-world scenarios.

Table. 4 Comparison table for speaker independent speech recognition implemented on mixed corpora

Paper	Database	Emotions	Feature extraction	Approaches	Recognition accuracy
2019 [20]	EMOVO, Emo-DB, IEMOCAP, EPST, RAVDESS, SAVEE, TESS	Positive, negative, neutral	Mel filterbank coefficients	CNN, LSTM, CNN-LSTM	Average accuracy was 53.35
2021 [21]	SAVEE, Urdu, EMO-DB, EMOVO	Neutral, happy, surprise	MFCC, spectral, energy, pitch, Chroma	Ensemble learning	Highest accuracy was 63.26%
2022 [22]	IEMOCAP	Neural, angry, sad, happy	Magnitude, Phase	CNN with attention	Up to 57.58% weighted accuracy
2022 [23]	RAVDESS	Neutral, calm, happy, sad angry, fearful, surprise disgust	MFCC, Spectrogram, Chroma, centroid, Rolloff	DNN	73.95%
2023 [24]	IEMOCAP and MSP-IMPROV	Anger, sadness, happy, neutral	Spectrogram	CNN with attention bi-LSTM	Up to 70.27%
Present work	Mixed Corpus	Happiness, Sadness, Anger, Fear, Neutral, Disgust, Surprise	MFCC	CNN, LSTM, CNN-LSTM	**Up to 76.16%**

5. CONCLUSION AND FUTURE WORK

This work established a SER system that uses hybrid CNN–LSTM to recognise seven emotions in six corpora using an MFCC feature extraction method. The results show that the hybrid CNN–LSTM outperformed the others for gender-specific and individual corpora, whereas CNN performed slightly better in the gender-independent mixed corpus. Mixing corpora aims to make databases more realistic, as all three corpora were recorded in different environment and the participants for each database are from different geographic locations with varying accents. On the mixed corpus, up to 65.62% accuracy has been achieved. The accuracy of the male mixed corpus is lower than the accuracy of the female mixed corpus (i.e., 54.86% and 76.86%, respectively). Upgrading the strength of the emotion recognition system is still attainable by integrating databases. It is worth stating that the hybrid CNN–LSTM is useful, as it takes advantage of aspects of both classifiers. This helps the classier perform better than classifiers applied individually.

A comparison has also shown that this study has better recognition accuracy than other studies, but there is always scope for improvement. Further studies could involve experiments aiming to optimise deep learning architectures, and various other feature selection techniques could be applied to improve recognition accuracy.

This discovery could be extremely valuable in the development of an SER system for robots that deal with clients from all over the world. It will allow robots to connect with clients intelligently using emotional intelligence, which might have a significant impact on how people communicate with robots in the future. The researchers aim to explore more deep learning approaches to boost the accuracy even more. Future plans will apply the research in real-life circumstances to explore speech datasets where the audio is from a natural setting. Future researchers will have a tough time locating corpora for various languages in a natural setting, because there are few obtainable. Second, the selection of the algorithms that gives consistent performance in both natural and recorded situations across all languages.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor (2001), Emotion recognition in human-computer interaction, IEEE Signal processing magazine, 32-80, (Vol. 18). https://doi.org/10.1109/79.911197

[2] A. Revathi and C. Jeyalakshmi (2019), Emotions recognition: different sets of features and models, International Journal of Speech Technology, 473-482, (Vol. 22). https://doi.org/10.1007/s10772-018-9533-6

[3] S. Latif, R. Rana, S. Khalifa, R. Jurdak and J. Epps (2019), Direct modelling of speech emotion from raw speech, arXiv preprint arXiv, 1904.03833. https://doi.org/10.48550/arXiv.1904.03833

[4] B. W. Schuller (2018), Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends, Communications of the ACM, 90-99, (Vol. 61). https://doi.org/10.1145/3129340

[5] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan (2019), Speech recognition using deep neural networks: A systematic review, IEEE Access, 19143-19165, (Vol. 7). https://doi.org/10.1109/ACCESS.2019.2896880

[6] S. Haq and P. J. Jackson (2011), Multimodal emotion recognition, InMachine audition: principles, algorithms and systems, IGI Global, 398-423. DOI: 10.4018/978-1-61520-919-4.ch017

[7] M. Swain, A. Routray and P. Kabisatpathy (2018), Databases, features and classifiers for speech emotion recognition: a review, International Journal of Speech Technology, 93-120, (Vol. 21). https://doi.org/10.1007/s10772-018-9491-z

[8] M. R. Ahmed, S. Islam, A. M. Islam and S. Shatabda (2023), An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition, Expert Systems with Applications, 119633, (Vol. 218). https://doi.org/10.1016/j.eswa.2023.119633

[9] G. A. Prabhakar, B. Basel, A. Dutta and C. V. R. Rao (2023), Multichannel CNN-BLSTM Architecture for Speech Emotion Recognition System by Fusion of Magnitude and Phase Spectral Features using DCCA for Consumer Applications, IEEE Transactions on Consumer Electronics, 226-235, (Vol. 69). https://doi.org/10.1109/TCE.2023.3236972

[10] M. R. Falahzadeh, F. Farokhi, A. Harimi and R. Sabbaghi-Nadooshan (2023), Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition, Circuits, Systems, and Signal Processing, 449-492, (Vol. 42). https://doi.org/10.1007/s00034-022-02130-3

[11] C. Lu, Y. Zong, W. Zheng, Y. Li, C. Tang and B. W. Schuller (2022), Domain invariant feature learning for speaker-independent speech emotion recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2217-2230, (Vol. 30). https://doi.org/10.1109/TASLP.2022.3178232

[12] J. Ancilin and A. Milton (2021), Improved speech emotion recognition with Mel frequency magnitude coefficient, Applied Acoustics, 108046, (Vol. 179). https://doi.org/10.1016/j.apacoust.2021.108046

[13] U. Kumaran, S. R. Rammohan, S. M. Nagarajan and A. Prathik (2021), Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN, International Journal of Speech Technology, 303-314, (Vol. 24). https://doi.org/10.1007/s10772-020-09792-x

[14] C. Li, Z. Bao, L. Li and Z. Zhao (2020), Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition, Information Processing & Management, 102185, (Vol. 57). https://doi.org/10.1016/j.ipm.2019.102185

[15] S. T. Saste and S. M. Jagdale (2017), Emotion recognition from speech using MFCC and DWT for security system, In2017 international conference of electronics, communication and aerospace technology (ICECA), IEEE, 701-704. (Vol. 1). https://doi.org/10.1109/ICECA.2017.8203631

[16] S. Haq and P. Jackson (2023), Surrey Audio-Visual Expressed Emotion (SAVEE) Database, http://kahlan.eps.surrey.ac.uk/savee

[17] S. R. Livingstone and F. A. Russo (2018), The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, PloS one, 0196391, (Vol. 13). https://doi.org/10.1371/journal.pone.0196391

[18] K. Dupuis and M. K. Pichora-Fuller (2023), Toronto emotional speech set (TESS), University of Toronto, Psychology Department, 2010, https://tspace.library.utoronto.ca/handle/1807/24487

[19] M. J. Al-Dujaili and A. Ebrahimi-Moghadam (2023), Speech emotion recognition: a comprehensive survey, Wireless Personal Communications, 2525-2561, (Vol. 129). https://doi.org/10.1007/s11277-023-10244-3

[20] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger and G. Hofer (2019), Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition, In INTERSPEECH, 201656-1660. http://dx.doi.org/10.21437/Interspeech.2019-2753

[21] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan and T. R. Gadekallu (2021), Cross corpus multi-lingual speech emotion recognition using ensemble learning, Complex & Intelligent Systems, 1845-1854, (Vol. 7). https://doi.org/10.1007/s40747-020-00250-4

[22] L. Guo, L. Wang, J. Dang, E. S. Chng and S. Nakagawa (2022), Learning affective representations based on magnitude and dynamic relative phase information for speech emotion

recognition, Speech Communication, 118-127, (Vol. 136). https://doi.org/10.1016/j.specom.2021.11.005

[23] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, A. Alhadlaq and H. N. Lee (2022), Two-way feature extraction for speech emotion recognition using deep learning, Sensors, 2378, (Vol. 22). https://doi.org/10.3390/s22062378

[24] Z. T. Liu, M. T. Han, B. H. Wu and A. Rehmanmhk (2023), Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning, Applied Acoustics, 109178, (Vol. 202). https://doi.org/10.1016/j.apacoust.2022.109178

AUTHORS

Neha Prerna Tigga received her degree in Integrated Master of Computer Application from Birla Institute of Technology, Mesra, India. Currently she is a doctoral student at Birla Institute of Technology, Mesra, Ranchi, India. Her research focuses on the interdisciplinary application of machine learning to solve real-world problems in the field of psychology and psychiatry.

E-mail: tigganeha4@gmail.com

Shruti Garg holds PHD degree in Computer Engineering from BIT, Mesra, Ranchi. Her area of competence and interest includes psychological disorder prediction, Emotion Recognition, Soft Computing, Machine Learning. She is Assistant Professor in Department of Computer Science and Engineering at BIT, Mesra from last 14 years. At BIT, she has been teaching courses like Algorithm, Artificial Intelligence, Soft Computing, Optimization, Python and Machine learning for machine vision.

E-mail: gshruti@bitmesra.ac.in