ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240819123835

Original scientific paper
Received: 2024-08-19, Accepted: 2025-03-24

Study of Gender-Specific Emotion Expressivity in Speech Using MFCC and CNN

Mayuri Bapat, Shankar M. Mali*, Chandrashekhar Patil

Abstract: Analysis of sentiment is a pivotal component of natural language processing and has recently observed noteworthy evolutions. Still, the impact of gender on expressivity of the emotion remains an undiscovered area. The proposed work utilizes a broad range of over 12000 audio data samples from four different benchmarked datasets RAVDESS, CREMA-D, TESS, and SAVEE. Convolutional Neural Network (CNN) is used to identify and detect patterns and biases according to gender. The study found mixed-gender emotion accuracy at 84.26%, female emotion accuracy at 89.40%, and male emotion accuracy at 82.70%. The proposed work aims to demonstrate that the female voice is more expressive of emotion than the male voice by examining the difference in sentiment expression between genders. This research will enhance the insights of sentiment analysis and can be useful to contrivance industries ranging from customer service to human-system interaction.

Keywords: Convolution Neural Network (CNN); feature extraction; Mel-Frequency Cepstral Coefficients (MFCC); vocal dimorphism

1 INTRODUCTION

Emotion expressivity is the ability to communicate internal emotions through various channels, including voice, facial expressions, body language, and written text. It builds connections, encourages empathy, and improves understanding of human social interaction. Emotion expressivity improves sentiment analysis accuracy by capturing nuances in spoken content. Audio files play a significant role in analyzing and quantifying emotion expressivity through voice. Measurable characteristics that are derived from audio signals are known as audio features. These features offer important information for diversely different applications of audio processing like Gender recognition, music analysis, Speech Emotion Recognition (SER), etc. Labeled training data is used to categorize features using a classifier into predefined classes based on patterns. Researchers used different classifiers such as Support Vector Machines (SVM) [3, 6, 11, 13], Convolutional Neural Networks (CNNs) [1, 12], Recurrent Neural Networks (RNNs) [7], or Long Short-Term Memory networks (LSTMs) [15], Decision Trees and Random Forests [1], Gaussian Mixture Models (GMM) [16], k-Nearest Neighbors (k-NN)[17], Ensemble Methods (e.g., AdaBoost, Gradient Boosting), Hidden Markov Models (HMM) [14], Naive Bayes Classifiers [13], Extreme Gradient Boosting (XGBoost) [19], BERT and GRU [8] which works as decision-making models that learn and infer emotional content from speech signals based on extracted features. Different techniques used for audio signal processing and can be listed as Spectral Features [1], Pitch-related Features, Zero Crossing Rate (ZCR) [1], Mel-frequency Cepstral Coefficients (MFCC) [1, 6, 7, 11, 12, 15, 25], UniSpeech-SAT [2], Harmonic to Noise Ratio (HNR) [3], Prosodic Features [7], Fundamental Frequency (F0), Spectral Features [11], Harmonic-to-Noise, Energy Features, Formant Frequencies, Statistical Features, Tonal centroid features [11], Time-Frequency Representations (e.g. Spectrograms) [18] etc.

The proposed study aims to analyze the emotion expressivity in male and female voices using Convolutional Neural Networks. This analysis is conducted using four benchmark datasets, ensuring a robust and comprehensive evaluation of vocal expressivity across genders.

The research paper's enduring sections are organized as follows: An outline of pertinent research is given, Section 2 highlights the research overview and associated research gap, and Section 3 elaborates on methodology in detail. Section 4 presents the Findings and Interpretation. Section 5 provides the concluding remark along with the possible future research.

2 LITERATURE REVIEW

Every human voice has some specific attributes like pitch, tone, pace, linguistic patterns, and emotions of voice, Emotional components present in the voice define the mental health of the human along with decisions. This section gives a brief about the current scenario of the research work done in voice or audio sentiment analysis with attention to the use of CNN.

Rezapour et al. (2023) [1] - This study aimed to classify emotions in speech using audio features and machine learning models. Researchers used a limited dataset to extract audio features using models like one-dimensional convolutional neural network (conv1D) and random forest (RF). RF with feature selection achieved higher accuracy (69%), precision (72%), and recall (84%) for fear and calm. However, similar acoustic qualities caused the misclassification of anger as happiness, disgust as sadness, and fear as sadness. Atmaja et al. (2022) [2] - This study evaluates sentiment analysis and emotion recognition from speech using self-supervised learning models. The results indicate that two types of sentiment analysis produced the greatest outcomes. On the other hand, higher-class models performed poorly on tests involving sentiment analysis and emotion perception. Performance declines could have been caused by the dataset's imbalanced nature.

Hadhami et al. (2020) [3] - The paper presents an emotion recognition system using speech signals, utilizing a two-stage approach for feature extraction and classification. It extracts a 42-dimensional vector of audio features, uses Auto-Encoder for parameter selection, and uses Support

Vector Machines as a categorizer. Trials are led at the Ryerson Multimedia Laboratory. Khan et al. (2021) [4] - The research demonstrate the effectiveness of the new method in surpassing a baseline system for sentiment analysis on audio data. It also identifies opportunities for further advancement, emphasizing the need to address ASR challenges and explore the integration of pure speech features for improved speech-based sentiment detection.

Madanian et al. (2023) [5] present a systematic review of ML-based research in SER over the past decade, specifically focusing on data processing, feature selection/extraction, and classification steps. It provides a thorough analysis of problems and their remedies, such as low classification accuracy in Orator-unconstrained experiments. Furthermore, the review furnishes guidelines for SER evaluation, emphasizing common baselines and available metrics for experimentation. Selvaraj et al. (2016) [6] - This paper presents an emotion recognition method using the MFCC approach and Radial basis function network, focusing on gender classification using a support vector machine and pitch analysis, proving more accurate than the Back Propagation Network.

Yoon et al. (2018) [7] paper introduces a groundbreaking multimodal approach that effectively combines text and audio inputs, outperforming existing methods in emotion classification. Its success in mitigating misclassification issues related to the neutral class demonstrates its promise for advancing emotion recognition in speech analysis. Yong-hun-Lee [8] used BNC64 corpus data and three types of analyses: dictionary-based, GRU-based, and BERT-based. Despite similar sentiment word usage, women used more positive words. The BERT-based analysis revealed more gender differences, supporting previous studies and highlighting the development of gender differences in sentiment analysis methods. Marianne Latinus [9] Two ERP studies revealed neural correlates of all processing of speeches. Differences in pitch between female and male voices were observed at 87 ms, with N1 and P2 showing early gender effects. P2 differentiated male from female voices regardless of pitch, suggesting that voice gender processing involves two stages: initial pitch discrimination followed by a later, more precise gender identification at P2 latency. Chiara De Amicis [10] - The study analyzed 78,000 earnings conference calls between 2004 and 2018 to compare sentiment between female and male senior managers. Results show female executives use a more positive tone and less vagueness, indicating a linguistic feature. Financial analysts also exhibit gender bias, but the stock market reaction is influenced by the call's sentiment, not the executive's gender. Ref. [11] focuses on detecting gender from voice signals using techniques to identify relevant features. It analyses voice signal features using a dataset, studies machine learning models, and uses feature selection algorithms to improve classification models. Experimental results show that sub-features are crucial for enhancing performance efficiency. Deep learning and SVM models gave 99.97% best recall value, and 100% for feature extraction techniques using SVM.

A three-layer feature extraction technique is used by Uddin et al. [12] to classify gender and region from human voices. In the first layer, it extracts the fundamental frequency, spectral entropy, spectral flatness, and mode frequency; in the second layer, it uses MFCC to map the audio data; and in the third layer, it computes LPC. The method performs better on gender and region classification using a combined dataset.

Maghilnan et al. [13] research aims to analyze sentiment in speech transcripts that distinguish between speakers to recognize the emotions of distinct speakers. It examines various methods for speaker discrimination and sentiment analysis to identify efficient algorithms for this task.

From the literature, we found that there is a need for systematic study to demonstrate gender differences using speech sentiment study and the use of advanced machine learning and deep learning approaches are necessary [8]. It is also suggested that more trials are required to find the spatial-temporal discrimination [9]. Research should consider gender differences in vocal emotion recognition using fully naturalized speech datasets to enhance conservation validity. [14]

3 METHODOLOGIES

3.1 Datasets Used in Experiment

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset contains 1200 files for speech and 1200 files for songs. It is a collection of multimedia types that are preferably used to study speech emotion recognition: songs. Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) is an emotive multimodal actor dataset featuring 7,442 original clips from 91 actors. The Toronto Emotional Speech Set (TESS) analyses how age affects an individual's capacity for emotion recognition. It contains 200 neutral phrases imitating seven different emotions. Surrey Audio-Visual Expressed Emotion (SAVEE) is an emotion identification dataset consisting of 480 utterances performed by 4 male actors. To combine these datasets for analysis, preprocessing steps are essential to ensure uniformity across all sources. First, the audio files are converted to a consistent format, such as WAV, and normalized for sampling rate and bit depth. Then, emotion labels are standardized to create a unified taxonomy across datasets, aligning overlapping categories. Features such as MFCCs are extracted from the audio to maintain consistency in feature representation. Finally, the data sets are merged into a single data structure, ensuring a balanced distribution of the samples by emotions and gender to prevent bias during the project cycle, like training and evaluating the model.

Table 1 Metadata of different datasets

Sr. No.	Dataset	FA	MA	NF FV	NF MV	NE	NF	AR
1	RAVDESS	12	12	732	720	8	7356	-
2	CREMA-D	43	48	1849	2304	6	7442	20 and 74
3	TESS	2	-	2800	-	7	2800	26 and 64
4	SAVEE	-	4	-	480	7	480	27-31

3.2 Data Pre-processing

Audio pre-processing is an essential step in modifying unprocessed acoustic signals into a form appropriate for machine learning prototypes. The process begins with audio sampling, where the waveform is captured at a rate of 0.70. If required, resampling is performed to match the model's input requirements or reduce computational overhead, using a target rate. To improve signal quality, noise reduction is applied followed by silence trimming based on amplitude thresholds. These techniques help to remove irrelevant artifacts and ensure that only the relevant portions of the audio are processed, which is critical for tasks like voice expressivity analysis. The raw audio is then converted into feature representations using MFCC. MFCCs are chosen because they effectively capture the timbral qualities of speech, which is essential for recognizing subtle variations in vocal expressivity.

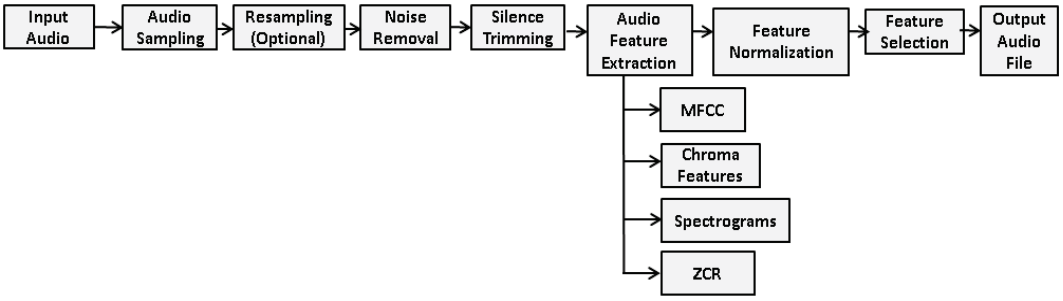

Figure 1 Audio Data Preprocessing Architecture

Table 2 Comparative Analysis of Techniques

Author	ED	MA	Dataset	FSM	AC	FSR
María et al. (2021)	7	Fully Convolutional Neural Network	EMODB, RAVDESS and TESS	Mel Spectrogram, MFCC	75.28 - RAVDESS 92.71 - EMODB, TESS -99.03	Evaluate auditory emotions broadly.
Eduyard et al. (2017)	6	CNN	Own Dataset	MFCC	71.33%	Explore cultural-language emotion identification.
Mu, Y. et al. [21]	4	CNN, BRNN Attention Model, LSTM	IEMOCAP	-	64.08 and 56.41 weighted and unweighted resp	Explore multimodal emotion identification regression.
Thomas, M. et al. [22]	-	RNN, LSTM	Malayalam datasets	-	80%	Deep learning enhances multilingual accuracy
Pavithra, P. et al. [23]	4	RNN	Own dataset	-	83%	personal assistant systems
Meenakshi, S. R. et al. [24]	-	RNN, LSTM, Word2Vec	PyAudio	-	-	Crowd Analysis
Huang, A. et al. [25]	8	SVM and HMM	RAVDESS TESS	MFCC, STFT	85%	Extend features and bidirectional LSTM training.
Jain, M. et al. [26]	4	SVM	Linguistic Data Consortium (LDC) and UGA database	MFCC, LPCC	85.085 %.	Enhance accuracy with MFCC, MEDC
Luo, Z. et al. [27]	5	Utterance-Based Parallel Neural Network	MOST	MFCC	68.72%	Integrate technologies for audio fusion
Cibau, N. E. et al. [28]	7	Autoencoder	EMO-DB	MFCC Prosodic features	70%	This can apply to digit recognition
Sahu, S. et al. [29]	4	Autoencoder SVM	IEMOCAP	Spectral, Prosody, and Energy features	58.38	Examine emotive speech in low-dimensional coding
Patel, N. et al. [30]	7	SVM, Decision tree classifier, CNN, Autoencoder	RAVDESS TESS	MFCC	96%	Replace decision trees with LSTMs/CNNs
Latif, S. et al. [31]	4	adversarial auto encoder	IEMOCAP MSP-IMPRO Librispeech	generative adversarial models	65.1	Integration of reinforcement learning expected
Hajarolasvadi, N. & Demirel, H. [32]	6	3D CNN	RML, SAVEE and eNTERFACE'05	MFCC	81.05% for SAVEE	Explore 3D architecture comparisons, data augmentation.

ED - Emotion Detected, MA - ML Algorithms, FSM - Feature Selection Method, AC - Accuracy, FSR - Further Scope of Research

These features are normalized using StandardScaler to ensure they share the same scale and distribution, which helps improve model convergence and performance. Data augmentation techniques, such as pitch shifting (with a pitch factor of 0.8), time stretching (with a factor of 0.75), and noise addition (with a sample rate of 0.70), are employed to expand the dataset. These augmentations are critical for simulating variations in voice due to factors like pitch, speed, and environmental noise. The use of data augmentation helped to address potential imbalanced classes in the dataset, ensuring

that the model learns to generalize across a broader spectrum of expressivity patterns. For example, by artificially altering the pitch and speed, the model can better recognize both male and female vocal expressivity across different conditions. Furthermore, feature determination methods are used to identify the most relevant characteristics that meaningfully affect the expressivity analysis. This ensures that the model does not overfit and focuses on the most important characteristics for classification. Finally, the pre-processed audio data are now ready for the machine learning model for further classification. In the proposed method Convolutional Neural Networks (CNN), are used which are well suited for capturing hierarchical patterns in Spectro-temporal features like MFCCs. This model will be used to analyze male and female vocal expressions, recognizing emotional and expressive cues from the processed audio data. Fig. 1 shows the audio data preprocessing architecture.

3.3 Different Approaches Used to Analyze Audio Sentiments

Convolutional Neural Networks (CNN) [32], Recurrent Neural Networks (RNN) [22-24], a combination of Convolutional and Recurrent Neural Networks (CRNN) [21, 37], Support Vector Machines (SVM) [26] and autoencoders [28-30] are mainly used and suitable for audio/voice/speech sentiment analysis. The use of one of these algorithms may depend on the characteristics of the dataset selected for the experiment and available resources. CNN master in feature extraction. The possibility of data overfitting is controlled by RNN. SVM is used to classify the emotions. Autoencoders are mainly used in unsupervised learning to focus on operations like data compression, feature learning, and dimensionality reduction. Tab. 2 gives a comparative analysis of these techniques with the scope of future research.

3.4 Feature Extraction Using MFCC

Feature extraction in audio files involves transforming raw data into meaningful features for easier analysis and classification by machine learning models or signal processing algorithms. Three axes-time, amplitude, and frequency represent the three dimensions of the audio stream [34]. The working of MFCC is depicted in Fig. 2.

Figure 2 Steps to perform MFCC

Audio Framing consists of Segmentation and Windowing. Segmentation divides the audio signal into short overlapping frames using techniques like the Hamming window. After that, apply a window function to each frame to minimize artifacts at frame boundaries. Fast Fourier Transform (FFT) helps to represent the frequency domain and to obtain a power spectrum. The frequency domain audio segment was converted from the temporal domain to the spectral domain using FFT. Calculation of the power spectrum from the magnitude of the FFT output is done to obtain the power spectrum. A Mel Filterbank is used to divide an audio sample into distinct Mel frequency scale frequency bands. To estimate the human perception of voice intensity Logarithmic Transformation takes the log of the filter. To Extract the high and low frequency changes from the audio signal Discrete Cosine Transform (DCT) is used. Applying DCT to log filter bank to de-correlate and coefficients helps to identify the most pertinent information. Coefficient Calculation usually discards higher-frequency coefficients, keeping a subset of the generated DCT coefficients as the final MFCCs. Feature Vector Construction is the next step, and it consists of two subsets: Vectorisation and Dynamic Features (Optional). Each frame is represented as a feature vector after vectorization aids in the selection of MFCC coefficients. Dynamic Features are an optional step support that computes delta and delta-delta coefficients to record the acceleration and rate of change of MFCCs between frames. The third stage, the final presentation, comprises the generated MFCC feature set and creates a series of feature vectors, each of which represents a distinct audio signal time window. By retaining certain spectral properties and eliminating others, these MFCC feature extraction methods seek to effectively represent the audio content.

3.5 Data Preparation

Data preparation involves preprocessing the data necessary for solving the problem of multiclass classification. Initially, all essential libraries like Scikit-learn were imported for preprocessing, evaluation, and data splitting. Secondly, labels and feature extraction are done from female voice and male voice datasets. Extracted features and labels for each gender are stored in separate variables. Using one-hot encoding categorical labels are transformed into binary matrix representations suitable for classification tasks. After concatenating the data for both genders, an 80-20 split is used to separate the merged dataset into training and testing sets. This ensures that the model is trained on a diverse dataset and evaluated on unseen data. The code further splits the datasets by gender, maintaining the same 80-20 train-test ratio. This separation allows for individual evaluation of the model's performance on female and male voices. To apply standardization, StandardScaler is used on training and testing datasets. This is achieved by scaling the features to have unit variance and zero means. This step is essential for ensuring that the structures are on a similar scale, improving the convergence of the training process. After standardization, the feature arrays are reshaped to add an extra dimension, making them compatible with the input requirements of Convolutional Neural Networks (CNNs). Lastly, the final shape of the dataset shows that the prepared data is three-dimensional, and the third dimension represents the channel of a single feature.

CNN expects input data in the form of RGB Channels,i.e. multiple channels, hence reshaping of the data is critical for CNNs. After this step, the dataset is ready to be fed into the proposed machine-learning model for training and evaluation. To illustrate probability distribution around the mean of continuous random variable Gaussian Curve or the Bell Curve is used. It is given in Eq. (1).

$$Z = \frac{(X - \mu)}{\sigma}, \tag{1}$$

where X is a normal arbitrary variable, μ is the mean of the distribution, σ is the standard deviation which measures the spread or variability of the distribution and z is the z-score over variable X. In summary, the code meticulously prepares the data for training a CNN by extracting features, one-hot encoding labels, splitting datasets, standardizing features, and reshaping arrays. These steps ensure that the data is in the optimal format for training an effective model to analyze gender-driven variations in sentiment.

3.6 Model Design

Convolutional Neural Network (CNN) is mainly used and suitable for audio or video processing [24]. It includes a sequential model, convolution layers, pooling, dropout, flattened, and dense layers. It master in feature extraction. Features may include identifying emotional changes, speakers, and their recognition, for example, speech pattern classification. In the case of voice expressiveness analysis using CNN, the process starts with input audio. The input audio is transformed into a spectrogram, which is a 2D representation of time versus frequency. A CNN applies a convolution operation wherein filters slide over the spectrogram to extract critical features such as pitch, tone, and amplitude variations. This operation generates feature maps, where each output captures meaningful characteristics related to voice expressiveness. In the proposed work, we have used multiple layers of convolution. The output layer is denoted by l and is calculated as given in Eq. (2).

$$y^{(l)}[m] = f\sum_{k=0}^{k-1} x^{(l-1)}[m-k]\cdot h^{(l)}[k] + b^{(l)}. \tag{2}$$

Where, $x^{(l-1)}$ denotes input to the lth layer, $h^{(l)}[k]$ are Learnable kernel weights for layer l, $b^{(l)}$ represents Learnable bias term for the layer l and $f(.)$ is a Non-linear activation function.

Pooling operations such as average pooling are applied to decrease the dimensionality of these feature maps while retaining the greatly significant characteristics and are continued by subsequent convolutional layers of 128 filters by wrapping up additional average pooling layers. The main task of the average pooling layer is to decrease the latitude dimensions, smooth the feature maps, and help embrace global context, which is crucial for accurate emotion detection. In the first layer of the model kernel size of 6 with 256 filters is used, then to dwindle feature map down sampling while maintaining important data average pooling is used. Subsequently, a convolutional layer with 64 filters is applied, followed by an average pooling layer that enhances key features by choosing the highest values from each pooling region.

To prevent overfitting by randomly omitting units during training, a dropout layer with a 20% dropout rate is incorporated, thus improving the model's robustness. After this, 3D feature maps are flattened into 1D vectors as a fact for fully connected condensed layers. A dense layer with 32 units combines the features nonlinearly, followed by another dropout layer with a 30% dropout frequency to decrease the overfitting of the model. The output layer, consisting of 8 units with a SoftMax activation function, provides a distribution of probabilities across the eight emotion groups, allowing multi-class organization. The Adam optimizer, well-known for its efficiency in the training of machine learning models, is used to build the model, as well as the category cross-entropy loss function. It is appropriate for solving problems involving multiple class categorizations. The accuracy metric is used to examine the performance of the model. Lastly, the features extracted are processed through wholly connected layers, which analyze and classify the features to determine specific voice expressiveness attributes. This structured approach makes CNN able to effectively model and interpret the nuances of voice expressiveness from audio data. Fig. 3 shows the CNN model architecture implemented. SVG-NN is used to create the proposed architecture.

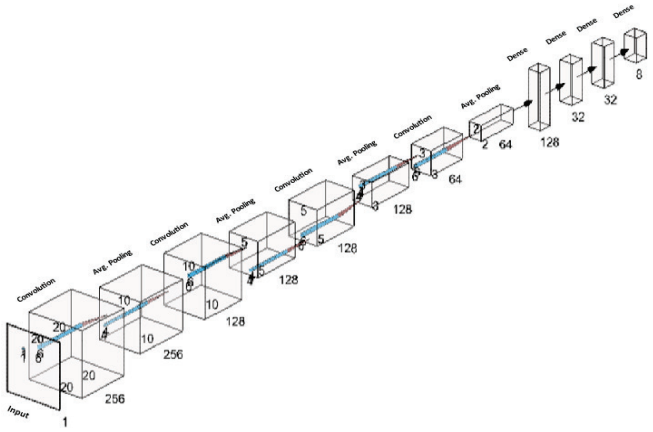

Figure 3 CNN Model Architecture

4 PERFORMANCE EVALUATION

The training features and matching labels are the inputs of the training and testing evaluation method, which is used to determine a proposed model's performance on two distinct datasets, i.e. training and testing. The results of this evaluation are stored in the score variable, which contains various evaluation metrics, such as accuracy. The print function is used to print the training and testing accuracy of the model up to two decimal places. To assess the model performance testing data is used in the second part of the model. Accuracy values give insights into the effectiveness of the final model in the prediction of male, female, and mixed-gender accuracy on training and testing data. Better

performance of the model is indicated with higher accuracy values. The generalization ability of the model is evaluated and provides valuable information about its effectiveness in real-world scenarios.

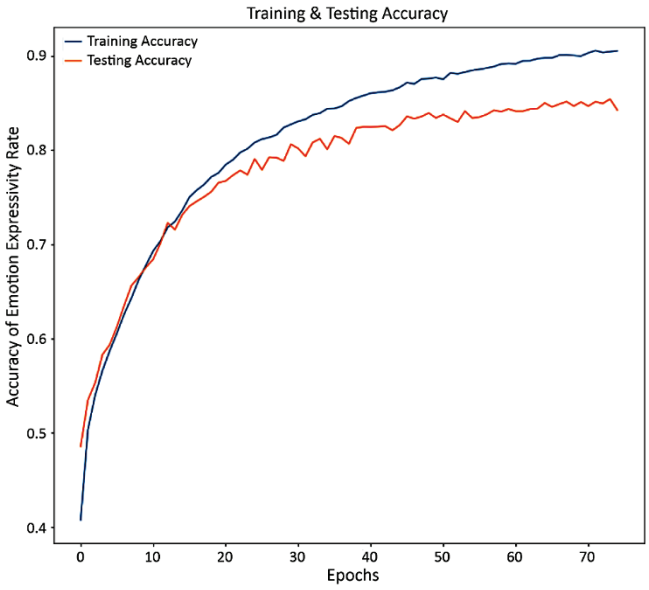

Figure 4 Accuracy across Gender

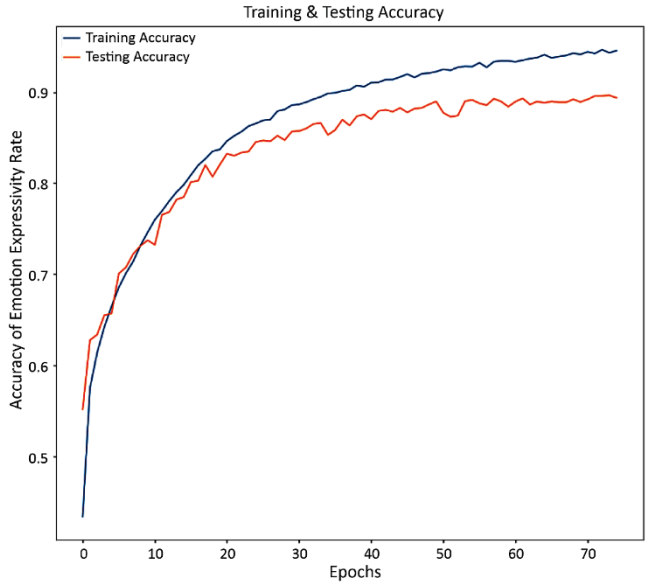

Figure 5 Female Voice Precision

The accurate emotion expressivity rate of mixed-gender, female, and male is presented in Figs. 4, 5, and 6 respectively. Distinct variations in accuracy across these categories are revealed by our analysis. It also highlights the impact of gender-specific characteristics on performance evaluation. Fig. 4 depicts the overall accuracy of emotion expressivity rate when analyzing a dataset that includes both male and female voices. The accuracy achieved in this mixed-gender scenario is 84.26%. This result provides a baseline understanding of the system's performance when gender-specific variations are not explicitly accounted for. Fig. 5 focuses on the accuracy of emotion expressivity rate specifically for female voices. The system gets an accuracy rate of 89.40% when analyzing female speech. This higher accuracy suggests that the emotional expressions in female voices are more consistently captured by the sentiment analysis model, indicating a better alignment between the attributes mined by the Mel-Frequency Cepstral Coefficients (MFCC) and the emotional content in female speech. Fig. 6 shows the accuracy of the emotion expressivity rate for male voices, which stands at 82.70%.

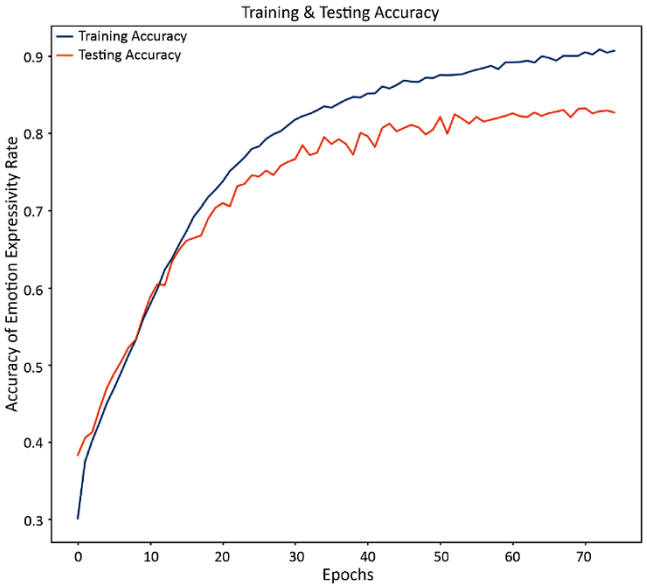

Figure 6 Male Voice Precision

From the results depicted in Figs. 4, 5, and 6, it is evident that female voices are more accurately analyzed for emotions, with a notable accuracy rate of 89.40%. These findings suggest that female voices may exhibit more pronounced or detectable emotional cues that align well with the features captured by MFCC. The disparity in accuracy rates underscores the importance of considering gender-specific characteristics in sentiment analysis.

A detailed comparative study of other research work focussed on emotion recognition of male, female, and mixed-gender is given in Tab. 3. Study shows that researchers have used different benchmark datasets like RAVDESS, EMODB, SAVEE, CASIA, etc. or some have developed their own datasets for research work. Different machine learning techniques are employed on varying sizes of voice samples. In the proposed work 12460 voice samples are used. The proposed work has a favorable outcome as CNN with average pooling is used on the integration of four datasets. It recognizes a broad spectrum of emotions and demonstrates its versatility and comprehensive approach. The proposed methodology results in the identification of expressiveness in a female voice in comparison with a male voice and from the obtained accuracy it is proved. As compared to other work proposed model provides competitive and balanced performance, making it a reliable solution for emotion recognition tasks across different gender-specific and mixed-gender datasets.

Table 3 Comparative analysis of various research studies on emotion recognition rates across male, female, and mixed-gender

Related Work	Dataset	Model	Voice Samples Used	Male Sentiment Credit Rate	Female Sentiment Credit Rate	Mix Gendered Credit Rate	Work Done
Nasaruddin, N. et al. [20]	Dataset from Kaggle	CNN with ResNet50 and ResNet101	3000	-	-	ResNet50: 99.67 ResNet101:99.82	Gender classification and detection
Kanwal, S. et al. [33]	RAVDESS EMO-DB SAVEE	SVM	1440	75.49 86.2 -	91.12 88.3 -	82.59 89.6 77.7	Speaker-dependent and speaker-independent speech recognition
Sun, T. W. [35]	CASIA	CNN	1200	-	-	84.60	Speech recognition with and without gender information.
Madhu, M. et al. [36]	RAVDESS	SVM	1440	-	-	72.02	Gender Recognition Using Speech
Singh, V. et al. [37]	RAVDESS	CNN	1440	-	-	72.07	Gender Dependent Speech Emotion Recognition System
Bhukya, S. [38]	Own Dataset	K-means algorithm	400	78	84	58	Gender Difference Identification and Genderize Speech Recognition
Dewan Arpita, H. et al. [39]	corpus of Bengali conversations	CNN	3185			99.37	Gender Identification Using Bengali Speech
Proposed Work	RAVDESS, TESS, SAVEE and CREMA-D	CNN with Average Pool	12460	82.70	89.40	84.26	Gender-Based expressivity of speech

5 CONCLUSION AND FUTURE ENHANCEMENT

We have achieved the key objective of the venture which is to identify the difference between the expressiveness of emotions of males and females using convolutional neural networks with MFCC. To fulfill this requirement, we have used four benchmark datasets that include 8 different emotions. From the accuracy mentioned in the above section, it is clear that the female voice is more expressive than the male voice. Following are the valuable contributions of the proposed research work: 1) We have developed a method to separate male and female voice files. 2) Synthetic training samples are created by adding perturbations to the initial training set. 3) The generalized CNN model was developed to capture the expressivity of the male, female, and mixed-gender voices using pitch data of the voice. 4) From the accuracy of emotion expressivity rate we have proved that female voice is more expressive than male voice. Additionally, we would like to use the proposed method in a real-time system. One can use the transfer learning technique along with the exploration of pre-trained models on voice data for sentiment analysis to improve accuracy. The proposed research can be extended by integrating speech data with different strategies like facial expressions and text. Exploring multimodal deep learning models for sentiment analysis possibly will lead to more precise and robust results. The future research directions mentioned above would contribute to further advancements in this field and enable the development of more sophisticated sentiment analysis systems with practical applications in various domains.

6 REFERENCES

[1] Mashhadi, M. M. R., & Osei-Bonsu, K. (2023). Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *PLoS ONE, 18*(11), e0291500. https://doi.org/10.1371/journal.pone.0291500

[2] Atmaja, B. T., & Sasou, A. (2022). Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations. *Sensors, 22*(17), 6369. https://doi.org/10.3390/s22176369

[3] Aouani, H., & Ayed, Y. B. (2020). Speech Emotion Recognition with deep learning. *Procedia Computer Science, 176*, 251-260. https://doi.org/10.1016/j.procs.2020.08.027

[4] Khan, P. A., Sumanth, T., & Vardhan, K. V. (2021). Audio sentiment analysis. *International Journal of Creative Research Thoughts (IJCRT), 9*(5), e835-e837. https://ijcrt.org/papers/IJCRT2105531.pdf

[5] Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning — A systematic review. *Intelligent Systems with Applications, 20*, 200266. https://doi.org/10.1016/j.iswa.2023.200266

[6] Selvaraj, M., & Karthik, S. P. (2016). Human speech emotion recognition. *International Journal of Engineering and Technology (IJET), 8(1),* 311-323. https://www.researchgate.net/publication/299185942_Human_speech_emotion_recognition

[7] Yoon, S., Byun, S., & Jung, K. (2018). Multimodal Speech Emotion Recognition Using Audio and Text. *arXiv* [Cs.CL]. http://arxiv.org/abs/1810.04635 https://doi.org/10.1109/SLT.2018.8639583

[8] Lee, Y.-H., & Kim, J.-H. (2021). A sentiment analysis of men's and women's speech in the BNC64. In K. Hu, J.-B. Kim, C. Zong, & E. Chersoni (Eds.), *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation* (pp. 603-610). https://aclanthology.org/2021.paclic-1.63

[9] Latinus, M., & Taylor, M. J. (2011). Discriminating Male and Female Voices: Differentiating Pitch and Gender. *Brain Topography*, 25(2), 194-204. https://doi.org/10.1007/s10548-011-0207-9

[10] De Amicis, C., Falconieri, S., & Tastan, M. (2021). Sentiment analysis and gender differences in earnings conference calls. *Journal of Corporate Finance, 71*, 101809. https://doi.org/10.1016/j.jcorpfin.2020.101809

[11] Alkhawaldeh, R. S. (2019). DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network. *Scientific Programming*, *2019*, 1-12. https://doi.org/10.1155/2019/7213717

[12] Uddin, M. A., Pathan, R. K., Hossain, M. S., & Biswas, M. (2021). Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN. *Journal of Information and Telecommunication, 6*(1), 27-42. https://doi.org/10.1080/24751839.2021.1983318

[13] Maghilnan S., & Rajesh Kumar, M. (2018). Sentiment analysis on speaker-specific speech data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1802.06209

[14] Lausen, A., & Schacht, A. (2018). Gender Differences in the Recognition of Vocal Emotions. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00882

[15] Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov Model-based Speech Emotion Recognition. ICASSP, IEEE *International Conference on Acoustics, Speech, and Signal Processing – Proceedings (ICASSP)*. 2. 401-404. https://doi.org/10.1109/ICME.2003.1220939

[16] Patel, P., Chaudhari, A. A., Pund, M. A., & Deshmukh, D. H. (2017). Speech Emotion Recognition System Using Gaussian Mixture Model and Improvement Proposed via Boosted GMM. *IRA-International Journal of Technology & Engineering, 7*(2 (S)), 56. https://doi.org/10.21013/jte.ICSESD201706

[17] Lanjewar, R. B., Mathurkar, S., & Patel, N. (2015). Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and k-Nearest Neighbor (k-NN) Techniques. *Procedia Computer Science, 49*, 50-57. https://doi.org/10.1016/j.procs.2015.04.226

[18] Luitel, S., & Anwar, M. (2022). Audio Sentiment Analysis using Spectrogram and Bag-of-Visual-Words. *The IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, San Diego, CA, USA, 200-205. https://doi.org/10.1109/IRI54793.2022.00052

[19] García-Ordás, M. T., Alaiz-Moretón, H., Benítez-Andrades, J. A., García-Rodríguez, I., García-Olalla, O., & Benavides, C. (2021). Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network. *Biomedical Signal Processing and Control, 69*, 102946. https://doi.org/10.1016/j.bspc.2021.102946

[20] Nasaruddin, N., Pratama Tresma, M. A. P., Muchamad, M. K., & Fuadi, Z. (2024). Voice frequency-based gender classification using convolutional neural network for smart home. *IEEE Access, 12*, 104190-104203. https://doi.org/10.1109/ACCESS.2024.3434547

[21] Mu, Y., Gómez, L. a. H., Montes, A. C., Martínez, C. A., Wang, X., & Gao, H. (2017). Speech Emotion Recognition Using Convolutional-Recurrent Neural Networks with Attention Model. *DEStech Transactions on Computer Science and Engineering*. https://doi.org/10.12783/dtcse/cii2017/17273

[22] Thomas, M., & Latha, C. A. (2018). Sentimental analysis using recurrent neural network. *International Journal of Engineering & Technology, 7*(2.27), 88. https://doi.org/10.14419/ijet.v7i2.27.12635

[23] Pavithra, P., Priya, N., & Naveenkumar, E. (2022). Recurrent Neural Network Based Speech Emotion Detection using Deep Learning. *Journal of Science Technology and Research, 3*(1).

[24] Meenakshi, S. R., Kumar, S. D., Rajasekar, D., & Prasad, S. S. (2020, August). Sentiment analysis using recurrent neural networks. In *National Conference on Recent Advancements in Communication, 7*(08).

[25] Huang, A., & Bao, P. (2019). Human vocal sentiment analysis. *arXiv*. https://arxiv.org/abs/1905.08632

[26] Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using a support vector machine. *arXiv*. https://arxiv.org/abs/2002.07590.

[27] Luo, Z., Xu, H., & Chen, F. (2019, January). Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural networks. In *AffCon@AAAI* (pp. 80-87). https://doi.org/10.29007/7mhj

[28] Cibau, N. E., Albornoz, E. M., & Rufiner, H. L. (2013). Speech emotion recognition using a deep autoencoder. *Anales de la XV Reunion de Procesamiento de la Informacion y Control*, *16*, 934-939.

[29] Sahu, S., Gupta, R., Sivaraman, G., AbdAlmageed, W., & Espy-Wilson, C. (2018). Adversarial auto-encoders for speech-based emotion recognition. *arXiv preprint arXiv:1806.02146*. https://arxiv.org/abs/1806.02146 https://doi.org/10.21437/Interspeech.2017-1421

[30] Patel, N., Patel, S., & Mankad, S. H. (2022). Impact of autoencoder-based compact representation on emotion detection from audio. *Journal of Ambient Intelligence and Humanized Computing*, *13*(2), 867-885. https://doi.org/10.1007/s12652-021-02979-3

[31] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., & Schuller, B. W. (2020). Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Transactions on Affective Computing*, *13*(2), 992-1004. https://doi.org/10.1109/TAFFC.2020.2983669

[32] Hajarolasvadi, N., & Demirel, H. (2019). 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*, *21*(5), 479. https://doi.org/10.3390/e21050479

[33] Kanwal, S., & Asghar, S. (2021). Speech emotion recognition using clustering-based GA-optimized feature set. *IEEE Access*, *9*, 125830-125842. https://doi.org/10.1109/ACCESS.2021.3111659

[34] Zhang, L. M., Li, Y., Zhang, Y. T., Ng, G. W., Leau, Y. B., & Yan, H. (2023). A deep learning method using gender-specific features for emotion recognition. *Sensors*, *23*(3), 1355. https://doi.org/10.3390/s23031355

[35] Sun, T. W. (2020). End-to-end speech emotion recognition with gender information. *IEEE Access*, *8*, 152423-152438. https://doi.org/10.1109/ACCESS.2020.3017462

[36] Nashipudimath, M. M., Pillai, P., Subramanian, A., Nair, V., & Khalife, S. (2021). Voice feature extraction for gender and emotion recognition. In *ITM Web of Conferences* (Vol. 40, p. 03008). EDP Sciences. https://doi.org/10.1051/itmconf/20214003008

[37] Singh, V., & Prasad, S. (2023). Speech emotion recognition system using gender-dependent convolution neural network. *Procedia Computer Science*, *218*, 2533-2540. https://doi.org/10.1016/j.procs.2023.01.227

[38] Bhukya, S. (2018). Effect of gender on improving speech recognition system. *International Journal of Computer Applications*, *179*(14), 22-30. https://doi.org/10.5120/ijca2018916200

[39] Dewan Arpita, H., Al Ryan, A., Fahad Hossain, M., Sadekur Rahman, M., Sajjad, M., & Noor Islam Prova, N. (2025). Exploring Bengali speech for gender classification: machine learning and deep learning approaches. *Bulletin of Electrical Engineering and Informatics, 14*(1), 328-337. https://doi.org/10.11591/eei.v14i1.8146

Authors' contacts:

Mayuri Bapat
Department of Computer Science and Application,
Dr. Vishwanath Karad MIT World Peace University,
Pune, Maharashtra, India 411038
mmbapat@mitacsc.ac.in

Shankar M. Mali
(Corresponding author)
Department of Computer Science and Application,
Dr. Vishwanath Karad MIT World Peace University,
Pune, Maharashtra, India 411038
shankar.mali@mitwpu.edu.in

Chandrashekhar Patil
Department of Computer Science and Application,
Dr. Vishwanath Karad MIT World Peace University,
Pune, Maharashtra, India 411038
chpatil.mca@gmail.com