JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# ADVANCING SPEECH EMOTION RECOGNITION THROUGH HIERARCHICAL ANALYSIS AND 2D-CONVOLUTIONAL NEURAL NETWORKS

**Ravi Prabhat Sharma[1] and Ritu Kadyan[2]**

Department of CSE
Ganga Institute of Technology & Management, Haryana, India

**Abstract.** Speech emotion recognition is essential for understanding and responding to human emotions in human-computer interaction. In this study, a 2D convolutional neural network (CNN) is proposed as a novel method for speech emotion recognition. The objective is to accurately classify emotions from speech signals by leveraging hierarchical and spatial information in the spectrogram representation. The research begins with audio data pre-processing, including noise removal and normalization. Features are then extracted, specifically Mel-frequency cepstral coefficients (MFCCs), capturing spectral characteristics. These MFCC features are transformed into a 2D spectrogram representation. A 2D-CNN architecture is designed and trained to learn discriminative features from the spectrograms. The proposed model employs multiple convolutional layers with varying filter sizes to capture local and global patterns. Max-pooling layers down-sample feature maps, improving generalization and information extraction. Performance evaluation uses publicly available speech emotion datasets (RAVDESS, SAVEE, TESS, and CREMA-D), showcasing the effectiveness of the 2D-CNN model in emotion recognition. This research contributes a novel approach that demonstrates accurate emotion classification, outperforming traditional methods and enhancing emotional intelligence in machines.

**Keywords:** Speech emotion recognition, MFCC, 2D-CNN architecture, speech emotion datasets, CREMA-D, RAVDESS, SAVEE, TESS, emotion classification.

## 1 INTRODUCTION

The field of artificial intelligence (AI) has experienced significant growth and advancements across various product industry sectors. One area of AI research that has attracted considerable attention is affective computing, which focuses on utilizing computational approaches to understand, process, and generate human emotions [1]. Affective computing has found applications in diverse fields such as marketing, advertising, medicine, and psychotherapy [2]. Emotion detection, an important component of affective computing, facilitates the expansion of facial expression recognition, empathy generation, language understanding, and other functionalities aimed at mimicking human behaviour. Furthermore, the advancement of affective computing technology as a whole depends on emotion recognition. It involves identifying human feelings from different data sources including text, gestures, voices, behaviours and facial expressions [3]-[5].

Feature selection and model selection are 2 main difficulties researchers face when working on speech emotion recognition. It involves computational modelling of human speech to identify emotions, with applications in fields such as forensics, security, and biometrics [6]. Apart from lexical content, human speech conveys additional characteristics such as gender, age, emotion and language [6] [7]. Numerous characteristics of audio data require translation into a computational representation, which is frequently represented as an array graph. The four categories into which the extraction of speech waveform data is often separated are Teager energy operator characteristics, Spectral, Prosodic, and audio quality.[8]. Among these features, As a cepstral domain characteristic, MFCC is frequently employed in speech emotion research.[12]. Selection of extracted audio features significantly influences the accuracy of emotion prediction, with MFCC demonstrating improved performance in previous studies [10], [13]-[15]. Segmentation techniques can be applied to the speech data to enhance the utility of these features. Hence, the accuracy of the emotion prediction. For instance, frame-based segmentation for segments with singular emotional speech was employed by Yeh et al. [16], while voiced segmentation proved more effective for longer speeches. Deep learning models have shown effectiveness in learning and recognizing emotions by leveraging their ability to identify patterns in visual data, such as facial expressions and voice structures [1]. These models offer advantages for emotion recognition, including automated feature selection that captures crucial emotion attributes, particularly in audio data [17]. Various classification models, including RNN or Recurrent Neural Networks, CNN or Convolutional Neural Networks, and SVM or Support Vector

Machines, have been applied in emotion recognition [10], [14], [17]. CNNs, known for their excellence in processing images, show promise in speech emotion recognition as they can represent extracted audio features as images. Previous research indicates that CNNs outperform RNNs and SVMs in speech emotion recognition tasks [14], [18]. Therefore, CNN is the platform we use for our speech emotion recognition model. A common An input layer, a convolutional layer, a pooling layer, a fully connected layer, and an output layer are the several layers that make up the CNN architecture.[19]. Convolution and pooling processes are crucial operations in CNNs, with filters used in the convolution process to extract feature maps containing essential information, while pooling process downsamples the feature maps, reducing dimensionality [20]. In previous experiments, CNNs showed remarkable accuracy in learning emotions for speech emotion recognition.[13]-[15], [22]. Researchers use a variety of audio datasets that range in language, duration, and expressed emotions. Many of these datasets are labelled with emotion categories and include the SAVEE, CASIA, IEMOCAP, RAVDESS, TESS and EMO [9], [11], [14], [15], [22], [23]. Combination of the datasets has been explored to enhance model performance, as demonstrated by de Pinto et al. [15]. By combining the RAVDESS and TESS datasets, they addressed the issue of overfitting observed in their previous work, resulting in improved model performance. Overfitting occurs when a network fails to effectively learn from data due to various factors such as noise in the training dataand ineffective architecture [24]. Combining datasets provides a larger quantity of data, thereby enhancing model performance by increasing the number of training instances. Using a 2D-CNN that integrates MFCC characteristics, audio segmentation, and bigger dataset, we present a unique method for recognising speech emotions in this study. We combine the CREMA, RAVDESS, TESS, and SAVEE datasets, which share similar pitch-based emotions and audio lengths, encompassing 7 common emotions: sad, neutral, surprise, happy, disgust, angry, and fear. Our classification model utilizes a 2D-CNN architecture and incorporates MFCC as an additional dimension, contributing to the accuracy of emotion prediction in speech emotion recognition.

## 2 Data Pre-Processing

In our research, we performed pre-processing on the input data using two common techniques:

a. Standard Scaler and
b. One-Hot Encoder.

Standard-Scaler is a widely used technique for feature scaling in machine learning tasks. It is particularly useful when the input features have different scales or units. By subtracting off the mean and dividing by the standard deviation of each feature, we applied Standard-Scaler to standardise the features. By reducing the features to a same scale and ensuring that they have zero mean and unit variance, this procedure makes it easier for machine learning algorithms to learn from the data. One-Hot Encoder, on the other hand, is employed for categorical feature encoding. It is used when we have categorical variables that need to be converted into a numerical representation suitable for machine learning algorithms. One-Hot Encoder creates binary dummy variables for each category of a categorical feature. A binary vector is used to represent each category, with a 1 representing that category and a 0 representing all other categories. The machine learning algorithm can successfully process categorical information thanks to this encoding.

By applying StandardScaler and OneHotEncoder preprocessing techniques to our input data, we ensure that the data is properly scaled and encoded, creating a suitable representation for training our machine learning models. These preprocessing methods standardise the features and properly encode categorical variables, which enhances the performance and accuracy of the models. After collecting the features and labels for both females and males, I divided the data into training and testing sets. I combined the data from both genders and split it into two parts: one for training and one for testing. As a result, the model was able to learn patterns in the training data and assess how well it performed using the testing data. I also repeated this process separately for females and males. To ensure fairness in comparing the results, I standardized the data. Standardization is like putting all the data on the same scale, so that they can be compared easily. It entails changing the values to have a 0 mean and a 1 standard deviation. This step enables the model to interact with the data in a useful way. After standardization, I made some modifications to the data's structure to match the requirements of the model. It involved expanding the dimensions of the data to include an additional aspect related to the audio features. This adjustment allows the model to process the data properly.

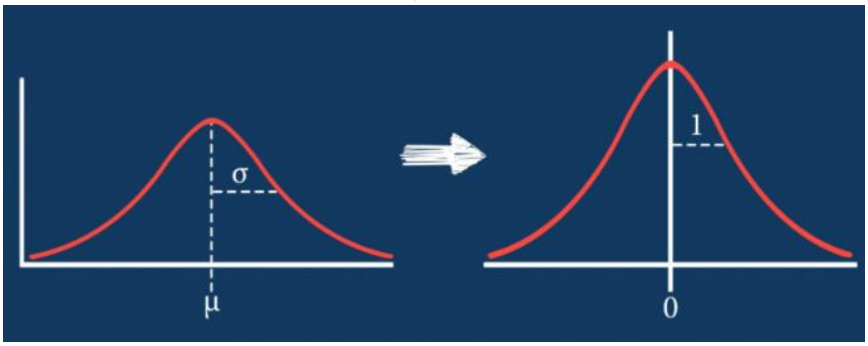


Figure 1 Z-Score Normalization (Standardization)

# 3 Methodology

## 3.1 Dataset Preparation:

For our research, we conducted a thorough search to identify available labelled speech audio datasets that meet the specific requirements for our experiments. These requirements include similarities in emotions, track duration, and compatibility with existing literature. By selecting datasets that fulfil these criteria, we aim to minimize potential accuracy reduction in deep learning algorithms [25]. After careful consideration, we chose three high-quality datasets: RAVDESS, SAVEE, CREMA and TESS. These datasets possess relevant similarities in terms of emotion representation and track duration. The RAVDESS dataset comprises audio speech and song recordings in English, featuring eight validated emotions [26]. It consists of a total of 1,440 speech data samples performed by 24 actors. The dataset includes expressions of neutral, calm, happiness, sadness, anger, fear, surprise, and disgust. SAVEE, on the other hand, is a recorded dataset that includes both visual and audio components. It was created by four researchers at the University of Surrey [2]. SAVEE contains 480 audio tracks in English, with each track expressing one of seven emotions: anger, disgust, fear, happiness, neutral, sadness, and pleasant surprise. The TESS consists audio recordings from two female performers. They express a range of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral [27]. TESS contains a total of 2,800 audio files in English. In addition to the RAVDESS, SAVEE, and TESS datasets mentioned earlier, another valuable dataset used in our research is the CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset). An extensive audio dataset with an emphasis on emotional speech is the CREMA-D dataset. It is made up of 91 performers, both male and female, who took part in controlled emotional outbursts and recorded their performances in high-quality audio. [28]. The dataset includes a wide range of emotions such as anger, disgust, fear, happiness, sadness, surprise, and neutral. With total of over 7,000 audio files, the CREMA-D dataset provides a rich collection of emotional speech data in the English language. Each audio file is annotated with emotional labels, ensuring that researchers have access to labelled data for emotion recognition tasks By incorporating these datasets into our research, we ensure access to a diverse range of labelled speech audio samples that cover various emotions and align with the objectives of our experiments. This selection process enhances the potential for robust and accurate analysis using deep learning algorithms.

## 3.2 Data Visualization and Analysis

First, by comparing two graphs to analyse the number of emotions separately for women and men. The emotions are arranged in a specific order: angry, calm, disgust, fear, happy, neutral, sad, and surprise. side by side, I gained insight into the distribution of emotions between women and men. This analysis provided valuable information about the emotional responses of the two genders and allowed me to observe possible differences or similarities in the expression of emotions.

Secondly, imagine we have recordings of people's voices when they express different emotions, like anger or joy. We try to understand how these emotions are reflected in the sound of their voices. The process I performed helped us analyse the recorded voice samples and understand how they sounded. It involved creating visual representations called waveplots, which displayed the changes in the voice over time. These waveplots were like pictures that captured the varying intensity of the voice as the recordings played. During the process, we had different recordings that portrayed various emotions, such as speaking with anger or singing in an angry tone. Using the available data, we generated waveplots for each emotion. These waveplots allowed us to identify distinctive patterns in the voice that corresponded to different emotions. To make the analysis even more immersive, we were able to listen to the recordings as well. By utilizing the process, we could play the audio for each emotion and experience how they sounded firsthand. In summary, the process enabled us to examine the sound patterns present in the recordings and gain insights into how emotions are expressed through our voices. It was like delving into how our feelings can be heard in the way we speak or sing.

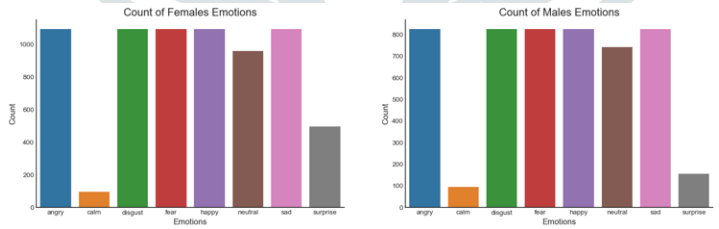


Figure 2 Visualization of Different Emotions for Women and Men

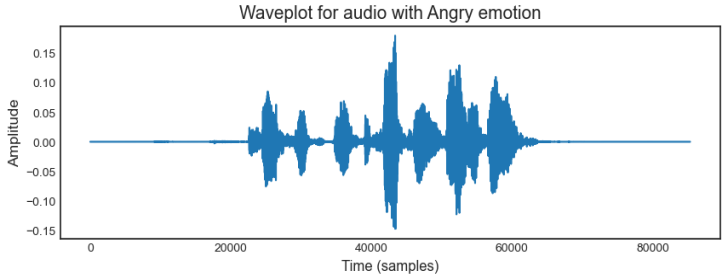


Figure 3 Waveplot for Audio with Angry Emotion

### 3.3 Data Augmentation:

Augmentation is a technique we use to enhance our training data by creating new synthetic samples. It involves introducing small changes or perturbations to our original set of training data. The goal is to make our model more robust and capable of generalizing well to unseen examples. The key idea behind data augmentation is to expose our model to a wider range of variations in the data, while still maintaining the same underlying label or emotion. This helps the model learn to handle different variations of the same emotion and improves its ability to accurately predict emotions in real-world scenarios. In the context of images, data augmentation techniques can include shifting the image, zooming in or out, or rotating it. However, in our case, since we are working with audio data, we apply different types of perturbations. These perturbations include adding noise, stretching or compressing the audio, and altering the pitch. By adding noise, we introduce random variations or disturbances to the audio signals. This helps the model become more robust to background noise or other environmental factors that may be present during real-world scenarios. Stretching or compressing the audio alters its duration, which helps the model learn to handle different speaking speeds or natural variations in speech patterns. Pitch shifting involves modifying the frequency or tone of the audio, allowing the model to recognize emotions even when the pitch of the voice varies. Overall, data augmentation is a valuable technique that expands our training data by creating new synthetic samples with slight modifications. This enables our model to learn from a more diverse range of examples and enhances its ability to generalize and accurately predict emotions in various real-world situations. Hence, to enhance the performance and generalization ability of our model, we applied data augmentation techniques to the audio data. These techniques involve introducing variations to the audio while preserving the original emotion label. We added noise to simulate real-world background noise and stretched or compressed the audio to account for different speaking speeds. We also shifted the audio samples to mimic timing variations and pitch-shifted the audio to capture changes in tone. Additionally, we increased and decreased the speed of the audio to account for faster or slower speech. By augmenting the data, we created diverse training samples that better represent real-world scenarios, improving our model's ability to recognize emotions in different conditions.

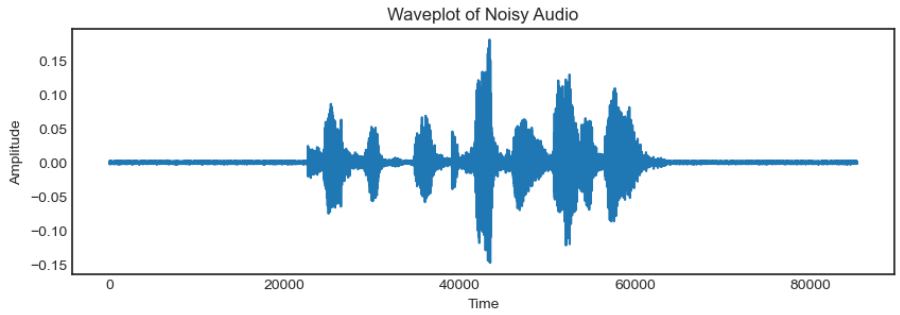


Figure 4 Waveplot for Noisy Emotion

### 3.4 Feature Extraction

In the field of speech recognition, feature extraction is a crucial step that converts audio samples into a format that can be easily understood and processed by models. The waveform of the audio alone may not provide sufficient information for classification purposes. Therefore, techniques like MFCCs are used to extract meaningful features from the audio waveform. MFCCs are widely used in audio signal processing because they capture important characteristics of the audio signal. They are derived through a series of computational steps that involve converting the audio waveform into a frequency domain representation, applying a filterbank to obtain the Mel-scale frequency components, and finally performing a Discrete Cosine Transform (DCT) to obtain the coefficients. By using MFCCs, unique characteristics of voice samples can be extracted, which can be used to differentiate between speakers or recognize different emotions. These coefficients provide a compact representation of the audio signal that captures relevant information while reducing complexity.

a. MFCCs are a widely used method for extraction of features in audio signal processing. They offer a concise representation of an audio signal's spectral envelope, capturing important characteristics of the signal's shape. The process of computing MFCCs involves several steps. First, the audio signal is divided into short frames, typically lasting around 20-40 milliseconds. Each frame undergoes a Fourier Transform to obtain its frequency spectrum. The mel scale, a perceptual scale of pitch that closely resembles how people hear sound, is then created using the frequencies. This conversion is done because human hearing is more sensitive to differences in lower frequencies than higher frequencies. Next, To accentuate differences in the lower frequencies and compress the dynamic range, the logarithm of the mel-scaled spectrum is used. This is followed by applying the Discrete Cosine Transform (DCT) to the logarithmic spectrum. The DCT reduces the dimensionality of the spectrum and decorrelates the coefficients, resulting in a smaller set of values that describe the spectral shape. Typically, around 10 to 20 MFCC coefficients are retained, representing the overall shape of the spectral envelope of the audio signal. These coefficients are effective in capturing key features related to speech and sound patterns, making them widely used in various applications such as speech recognition, speaker identification, and audio classification.

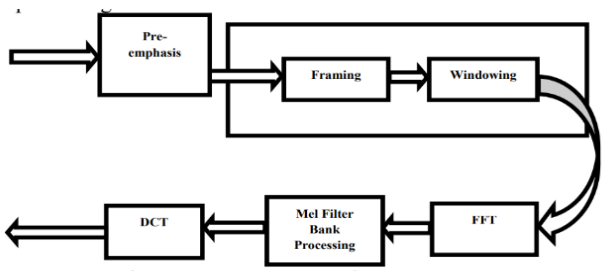


Figure 5 Feature Excerption Process

Thus, I implemented feature extraction from audio data using Mel-Frequency Cepstral Coefficients (MFCCs). I processed the audio samples by applying various transformations like adding noise, stretching, shifting, adjusting pitch, and changing the speed. These transformations help in augmenting the data and making it more diverse. Then, I extracted MFCC features from the augmented audio samples.

I organized the data based on gender, creating separate data frames for females and males. Each data frame contained the extracted features along with the corresponding emotion labels. If the data frames already existed, I loaded them from saved files instead. This way, I prepared the data for further analysis or machine learning tasks.

## 4 Proposed Approach

We begin by constructing a Convolutional Neural Network (CNN) model designed for classifying emotions. The CNN model employs multiple layers to process the data. The architecture and parameters are carefully selected to optimize its performance. For training the model, a pre-arranged dataset containing emotion-related information is loaded. The dataset is divided into segments of specific lengths. The data is transformed into MFCC arrays, representing the features used for classification. The data is then partitioned into training, test, and validation sets to evaluate the model's performance on unseen data. The CNN model is constructed with convolutional, pooling, dropout, and fully connected layers. Convolutional and pooling layers extract important patterns and features, while dropout layers prevent overfitting. A fully connected layer combines the extracted features for predictions. The categorical cross-entropy loss function and Adam optimizer are used in the model's construction. To compute loss and accuracy, performance evaluation is done using the validation set. To enhance training, a learning rate reduction technique is employed to adjust the learning rate when improvement stagnates. The batch size and number of training epochs are specified, and graphs are generated to visualize the changes in loss and accuracy during training. The model is constructed and evaluated separately for the total dataset, female dataset, and male dataset, with each model having its own architecture summary and accuracy score based on the respective test dataset. The training process consists of 75 epochs.

## 5 Results and Discussion

### 5.1 Evaluation

A comprehensive evaluation was conducted to assess the performance of various emotion recognition models based on gender. The objective was to determine the effectiveness of these models in accurately predicting emotions. The evaluation process involved analyzing the models' performance on training and testing datasets, focusing on their ability to generalize across genders. The analysis began with the examination of a "mixed-gender emotions" model. This model was trained and tested using a dataset that consisted of emotions expressed by individuals of diverse genders. During the training phase, the model achieved a high accuracy of 96.03%, indicating its proficiency in predicting emotions within the training data. To assess its generalization capability, model was tested on unseen data. Testing accuracy of 87.01% demonstrated that the model could effectively generalize and accurately predict emotions across different genders. Next, a specialized model designed specifically to recognize emotions expressed by females was evaluated. This model underwent training using a dataset comprising emotions expressed exclusively by females. The training accuracy of 99.86% reflected the model's exceptional performance in predicting emotions within the female training data. To assess its generalization ability, model was tested on unseen female data. Testing accuracy of 94.46% showcased the model's effectiveness in accurately predicting emotions in new instances within the female gender group.

Similarly, a model focused on recognizing emotions expressed by males was assessed. The model underwent training using a dataset that exclusively represented emotions expressed by males. It achieved a training accuracy of 97.33%, indicating its proficiency in predicting emotions within the male training data. During the testing phase, the model exhibited a testing accuracy of 86.45%, indicating a relatively more challenging task of generalizing to unseen male emotions compared to the specialized female model's performance on unseen female emotions. In conclusion, the evaluation revealed that the mixed-gender emotions model displayed reasonable performance on both the training and testing datasets, indicating its competence in predicting emotions across genders. However, the specialized models tailored to specific genders surpassed the mixed-gender model in terms of accuracy. The female emotions model demonstrated outstanding accuracy in recognizing and predicting emotions within the female gender group, while the male emotions model showed a slightly lower accuracy in generalizing to unseen male emotions. These findings suggest that developing separate models specifically tailored to individual genders can lead to improved accuracy in recognizing emotions within those gender groups. It is critical to remember that these results may differ when applied to various datasets or evaluated using alternative measures since they depend on the unique assessment metrics and dataset utilised in this study.

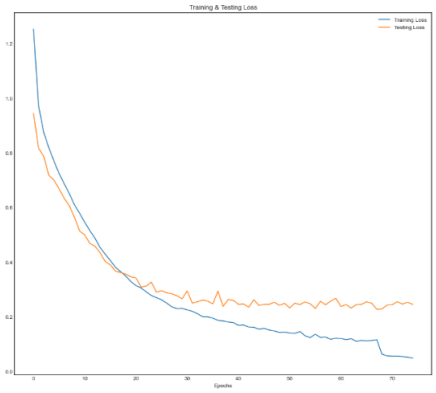


Figure 6 Mixed-gender emotions: Training and Validation Loss

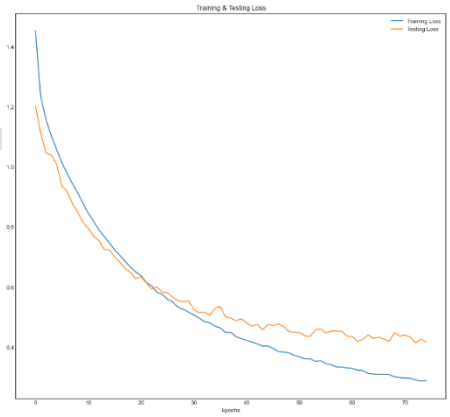


Figure 7 Accuracy of Training and Validation for Mixed-Gender Emotions

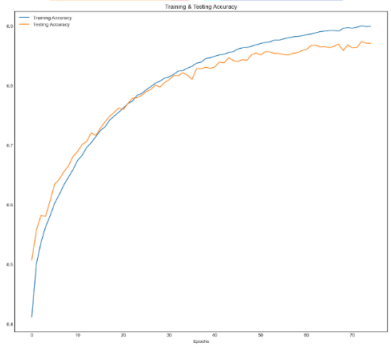


Figure 8 Training and Validation Loss for Female Emotions

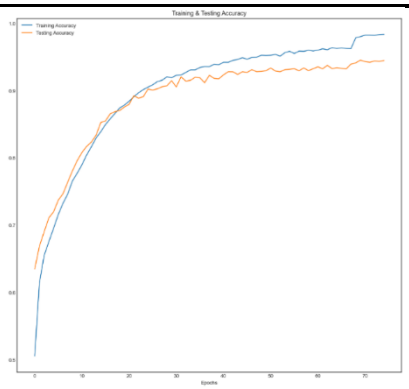


Figure 9 Accuracy in Training and Validation for Female Emotions

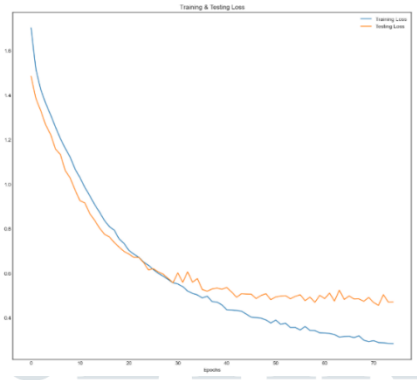


Figure 10 Male Emotions: Training and Validation loss

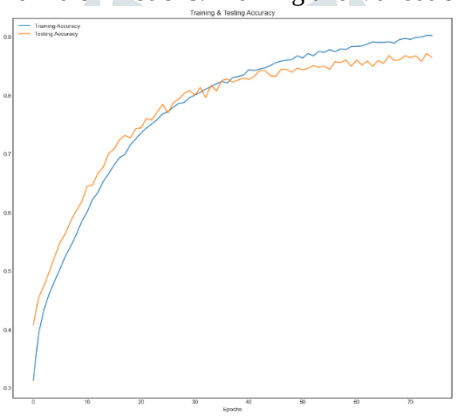


Figure 11 Male Emotions: Training and Validation Accuracy

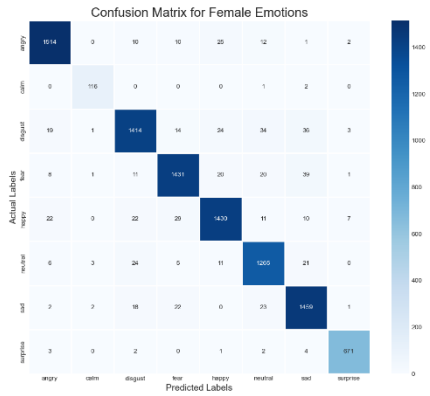


Figure 12 Female Emotions Confusion Matrix

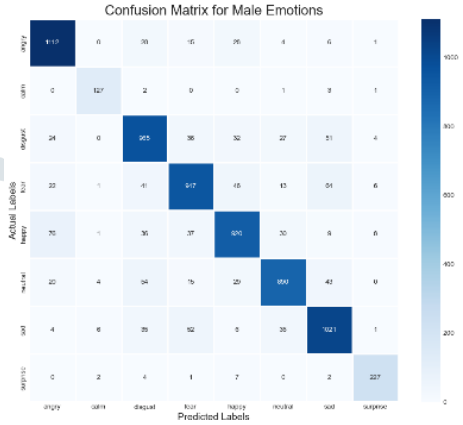


Figure 13 Male Emotions Confusion Matrix

## 6 Conclusion and Future Scope

In conclusion, our research focused on evaluating the performance of different emotion recognition models based on gender. We conducted a comprehensive analysis to determine the effectiveness of these models in accurately predicting emotions.

- Our findings revealed that the mixed-gender emotions model exhibited reasonably good performance, achieving a high training accuracy of 96.03%. When tested on unseen data, the model demonstrated a satisfactory accuracy of 87.01%, indicating its ability to generalize well across various genders.
- Furthermore, we investigated specialized models designed specifically for recognizing emotions expressed by females and males. The female emotions model showcased exceptional performance, achieving a remarkable training accuracy of 99.86%. It accurately predicted emotions within the female training data and achieved an impressive accuracy of 94.46% on unseen female data.
- Similarly, the male emotions model attained a notable training accuracy of 97.33%, indicating its proficiency in predicting emotions within the male training data. However, its testing accuracy of 86.45% suggested a relatively more challenging task of generalizing to unseen male emotions compared to the female model's performance.
- Based on our evaluation, it can be concluded that training separate models for specific genders yields better results in recognizing emotions within those gender groups. The specialized models for female and male emotions outperformed the mixed-gender model in terms of accuracy. This highlights the importance of considering gender-specific models in emotion recognition tasks.
- It is important to recognise that these findings are unique to the dataset and assessment criteria used in our study. Different datasets or alternative evaluation metrics may yield varying results. Future studies could explore additional factors such as cultural differences or age groups to further enhance emotion recognition models.
- Overall, our research contributes to the advancement of emotion recognition technology and emphasizes the significance of gender-specific models in accurately capturing and understanding emotions.

However, there are several avenues for future research to further enhance this field. One area is the exploration of intersectionality, considering how gender intersects with other social factors to influence emotion expression and recognition. Additionally, fine-grained emotion recognition could be pursued to capture nuanced emotions within specific gender groups. Transfer learning and domain adaptation techniques can be investigated to improve generalization capabilities, while real-time emotion recognition models can be developed for practical applications. Moreover, ethical considerations, including biases and privacy concerns, should be addressed, and strategies to mitigate bias and ensure fairness in predictions should be explored. By addressing these future research directions, the field of emotion recognition can advance both technically and ethically, leading to more accurate, inclusive, and responsible applications.

## 7 References


1. W. Dai, D. Han, Y. Dai, and D. Xu, "Emotion recognition and affective computing on vocal social media," Information & Management, vol. 52, no. 7, pp. 777-788, Nov. 2015, doi: 10.1016/j.im.2015.02.003.
2. F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "Recognizing emotion from speech based on age and gender using hierarchical models," Procedia Computer Science, vol. 151, pp. 37-44, 2019, doi: 10.1016/j.procs.2019.04.009.
3. J. Bhaskar, K. Sruthi, and P. Nedungadi, "Hybrid approach for emotion classification of audio conversation based on text and speech mining," Procedia Computer Science, vol. 46, pp. 635-643, 2015, doi: 10.1016/j.procs.2015.02.112.
4. M. T. P. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," Procedia Computer Science, vol. 108, pp. 1175-1184, 2017, doi: 10.1016/j.procs.2017.05.025.
5. D. Li, Y. Zhou, Z. Wang, and D. Gao, "Exploiting the potentialities of features for speech emotion recognition," Information Sciences, vol. 548, pp. 328-343, Feb. 2021, doi: 10.1016/j.ins.2020.09.047.
6. S. A, A. Thomas, and D. Mathew, "Study of MFCC and IHC feature extraction methods with probabilistic acoustic models for speaker biometric applications," Procedia Computer Science, vol. 143, pp. 267-276, 2018, doi: 10.1016/j.procs.2018.10.395.
7. P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech-A review," in Toward Robotic Socially Believable Behaving Systems - Volume I: Modeling Emotions, A. Esposito and L. C. Jain, Eds. Cham: Springer International Publishing, 2016, pp. 205-238.
8. M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Communication, vol. 116, pp. 56-76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.
9. J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," Applied Acoustics, vol. 179, Aug. 2021, doi: 10.1016/j.apacoust.2021.108046.
10. H. Aouani and Y. Ben Ayed, "Speech emotion recognition with deep learning," Procedia Computer Science, vol. 176, pp. 251-260, 2020, doi: 10.1016/j.procs.2020.08.027.
11. H. Murugan, "Speech emotion recognition using CNN," International Journal of Psychosocial Rehabilitation, vol. 24, no. 8, pp. 2408-2416, 2020, doi: 10.37200/IJPR/V24I8/PR280260.
12. G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," Applied Acoustics, vol. 158, 107020, Jan. 2020, doi: 10.1016/j.apacoust.2019.107020.
13. R. Y. Rumagit, G. Alexander, and I. F. Saputra, "Model comparison in speech emotion recognition for Indonesian language," Procedia Computer Science, vol. 179, pp. 789-797, 2021, doi: 10.1016/j.procs.2021.01.098
14. A. Bin Abdul Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional neural network (CNN) based speech-emotion recognition," in 2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON), 2019, pp. 122-125, doi: 10.1109/SPICSCON48833.2019.9065172.
15. M. G. de Pinto, M. Polignano, P. Lops, and G. Semeraro, "Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients," in 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), May 2020, pp. 1-5, doi: 10.1109/EAIS48028.2020.9122698.
16. J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, "Segment-based emotion recognition from continuous Mandarin Chinese speech," Computers in Human Behavior, vol. 27, no. 5, pp. 1545-1552, Sep. 2011, doi: 10.1016/j.chb.2010.10.027.
17. W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in 2016 AsiaPacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Dec. 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820699.
18. B. Zhang, C. Quan, and F. Ren, "Study on CNN in the recognition of emotion in audio and images," in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Jun. 2016, pp. 1-5, doi: 10.1109/ICIS.2016.7550778
19. K. O'Shea and R. Nash, "An introduction to convolutional neural networks," arXiv preprint arXiv,1511.08458, Nov. 2015.
20. R. Zhu, X. Tu, and J. X. Huang, "Deep learning on information retrieval and its applications," in Deep Learning for Data Analytics, H. Das, C. Pradhan, and N. Dey, Eds. Elsevier, 2020, pp. 125-153.
21. A. Pardamean and H. F. Pardede, "Tuned bidirectional encoder representations from transformers for fake news detection," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), vol. 22, no. 3, pp. 1667-1671, 2021, doi: 10.11591/ijeecs.v22.i3.pp1667-1671.
22. J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical Signal Processing and Control, vol. 47, pp. 312-323, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035
23. M. Gao, J. Dong, D. Zhou, X. Wei, and Q. Zhang, "Speech emotion recognition based on convolutional neural network and feature fusion," in 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 2019, pp. 1145-1150, doi: 10.1109/ISKE47853.2019.9170369.