Speech Emotion Recognition

Ebrahim Golriz
Data Mining mini-project

***Abstract*—This research focuses on developing a Speech Emotion Recognition (SER) system utilizing the RAVDESS dataset. Audio features were represented using 2D Mel-Frequency Cepstral Coefficients (MFCCs). The study compared the performance of Support Vector Machines (SVM) and Convolutional Neural Networks (CNNs) for SER. Results demonstrated that the CNN model outperformed the SVM, achieving higher accuracy. Furthermore, the study analyzed the performance of the CNN model in recognizing emotions and identifying gender independently.**

I. Introduction

As human communication comprises both verbal content and emotional expression, the ability to automatically detect and classify emotions from speech signals has profound implications for applications ranging from mental health monitoring to enhanced human-computer interfaces. The recognition of emotional states from speech presents unique challenges due to the complex interplay of acoustic features, speaker variability, and the subtle nature of emotional expression in vocal patterns.

In recent years, the advancement of machine learning techniques has revolutionized the approach to speech emotion recognition. Traditional methods relied heavily on hand-crafted features and conventional classifiers, while modern deep learning approaches promise more sophisticated feature learning and improved recognition accuracy. This project implements and compares Support Vector Machine (SVM) and Convolutional Neural Network (CNN), with a particular focus on the effectiveness of Mel-frequency cepstral coefficients (MFCCs) as acoustic features.

This project utilizes the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [1], a comprehensive dataset that provides controlled recordings of emotional expressions across multiple speakers of different genders. This dataset's structured nature, with its careful control of emotional intensity and technical recording quality, makes it particularly suitable for investigating the nuances of emotion recognition across gender boundaries.

II. Methodology

A. Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [1] dataset is a widely used resource for speech emotion recognition (SER) tasks. It comprises recordings of 24 professional actors (12 male and 12 female) expressing eight distinct emotions: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. Each recording includes both speech and song formats, providing a rich variety of vocal tones and emotional expressions. The audio files are recorded in a controlled environment with high-quality sampling, ensuring consistency and clarity. RAVDESS is particularly suited for SER tasks because of its balanced distribution of emotions, genders, and its multimodal nature, making it an excellent benchmark dataset for emotion recognition studies. This project focuses on only the audio portion of the speech part of this dataset.

The dataset was created following ethical guidelines. Human volunteers provided informed written consent before participation, and the recording methods were approved by the ethics committee of Ryerson University, Canada, in accordance with the Declaration of Helsinki. This ensures that the data is ethically sourced and suitable for research purposes [1].

B. Feature Extraction

Mel Frequency Cepstral Coefficients (MFCCs) [2] was chosen as the primary feature representation for the audio signals in this project. MFCCs are widely used in speech processing because they effectively capture the perceptual characteristics of sound, mimicking the human auditory system [3]. The 2D MFCC representations were extracted from the audio files to create time-frequency images suitable for further processing, an example of which is displayed on fig. 1. MFCCs was selected due to its simplicity, computational efficiency, and established performance in capturing features relevant to speech emotion recognition. While other features such as spectrograms or chroma might have been considered, MFCCs provided a sufficient balance between complexity and performance, making them ideal for the scope of this study.

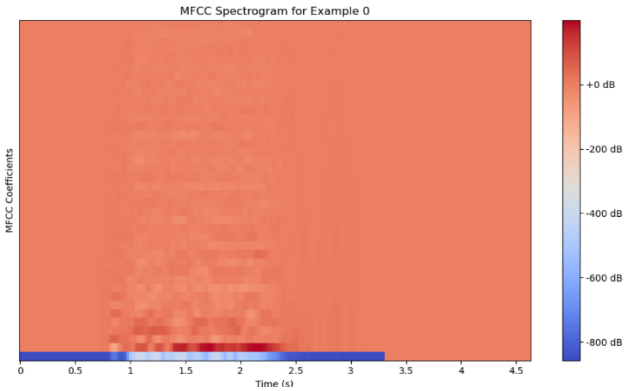

Fig. 1. Example of MFCCs of an audio track

C. Model Training

Two models were employed for the classification tasks: a Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). SVM was used as a baseline model due to its simplicity and effectiveness in high-dimensional feature spaces. It operates by finding an optimal hyperplane that separates different classes, making it a reliable starting point for classification tasks. On the other hand, the CNN model was designed to process the 2D MFCC images directly, which enables its ability to automatically learn spatial hierarchies and extract complex features. By stacking convolutional layers with pooling and activation functions, the CNN could capture nuanced patterns in the MFCC images, ultimately

achieving superior performance compared to the SVM. This highlighted the strength of deep learning approaches in SER tasks.

III. Results and Discussion

The performance of the SVM and CNN models was evaluated using the RAVDESS dataset. The classification metrics used include accuracy, precision, recall, and F1-score.

A. Model Performance

The SVM achieved an overall accuracy of 56%, with a weighted average F1-score of 0.53. While SVM performed reasonably well on some classes, such as "female_surprise" and "female_calm", it struggled with others, like "male_neutral" and "female_neutral", where it failed to classify correctly, which indicates a potential difficulty in distinguishing between neutral expressions. In contrast, the CNN achieved a higher accuracy of 60%, with a weighted average F1-score of 0.60, showing a clear improvement over SVM. However, a closer examination of the CNN's performance reveals that while it improved on the SVM's performance with "male_neutral" and "female_neutral", these classes still presented a challenge, which indicates nuanced differences between calm and neutral, and potentially across genders, required more sophisticated feature extraction than the current model offered. Overall, the CNN model demonstrated better consistency across most classes, particularly excelling in "female_surprise" and "female_calm". Both models were found to be particularly good at classifying "calm" and "surprise" emotions. This is likely because "surprise" has a very distinct sound (sudden changes in voice) and "calm" has a very consistent sound (steady and regular). These clear and contrasting sounds make it easier for the models to identify these emotions accurately. The complete and detailed classification report and confusion matrix of SVM and CNN models can be observed in Fig. 2, Fig. 3, Table I and Table II.

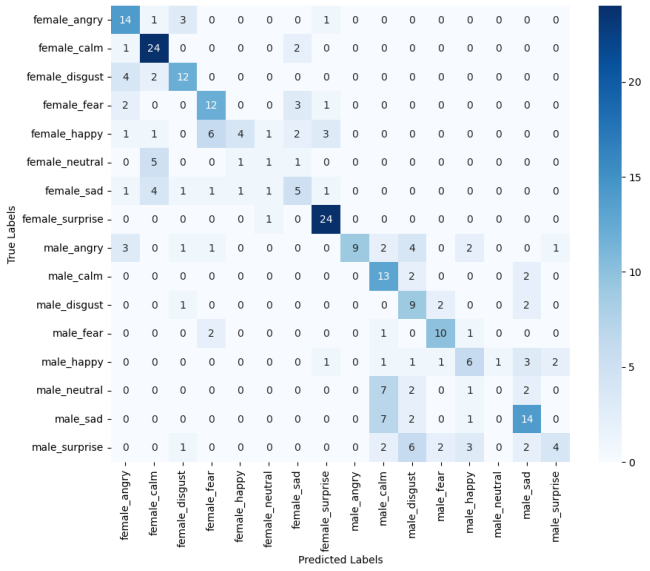

Fig. 2. Confusion matrix of SVM model

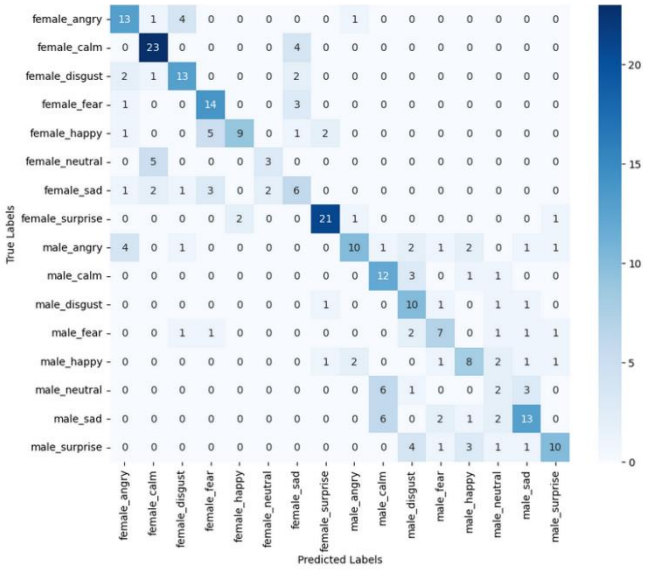

Fig. 3. Confusion matrix of CNN model

Table I. Classification Report of SVM Mode

Emotion	Precision	Recall	F1-Score	Support
female_angry	0.54	0.74	0.62	19
female_calm	0.65	0.89	0.75	27
female_disgust	0.63	0.67	0.65	18
female_fear	0.55	0.67	0.60	18
female_happy	0.67	0.22	0.33	18
female_neutral	0.25	0.12	0.17	8
female_sad	0.38	0.33	0.36	15
female_surprise	0.77	0.96	0.86	25
male_angry	1.00	0.39	0.56	23
male_calm	0.39	0.76	0.52	17
male_disgust	0.35	0.64	0.45	14
male_fear	0.67	0.71	0.69	14
male_happy	0.43	0.38	0.40	16
male_neutral	0.00	0.00	0.00	12
male_sad	0.56	0.58	0.57	24
male_surprise	0.57	0.20	0.30	20
accuracy			0.56	288
macro avg	0.53	0.52	0.49	288
weighted avg	0.57	0.56	0.53	288

Table II. Classification Report of CNN Model

Emotion	Precision	Recall	F1-Score	Support
female_angry	0.59	0.68	0.63	19
female_calm	0.72	0.85	0.78	27
female_disgust	0.65	0.72	0.68	18
female_fear	0.61	0.78	0.68	18
female_happy	0.82	0.50	0.62	18
female_neutral	0.60	0.38	0.46	8
female_sad	0.38	0.40	0.39	15
female_surprise	0.84	0.84	0.84	25
male_angry	0.71	0.43	0.54	23
male_calm	0.48	0.71	0.57	17
male_disgust	0.45	0.71	0.56	14
male_fear	0.54	0.50	0.52	14
male_happy	0.53	0.50	0.52	16
male_neutral	0.20	0.17	0.18	12
male_sad	0.62	0.54	0.58	24
male_surprise	0.71	0.50	0.59	20
accuracy			0.60	288
macro avg	0.59	0.58	0.57	288
weighted avg	0.62	0.60	0.60	288

B. *Gender Recognition*

The CNN model's results were further analyzed to group predictions by gender only. This analysis showed a highly accurate performance for gender recognition. The CNN achieved an overall accuracy of 96%, with an F1-score of 0.96 for both male and female categories. This indicates that the model was proficient at distinguishing between male and female voices, regardless of the emotions expressed. The detailed confusion matrix and classification report can be seen in Fig. 4 and Table III.

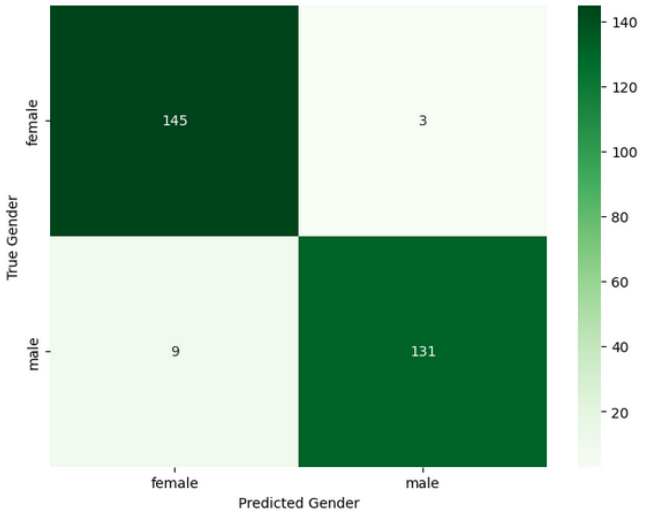

Fig. 4. Gender-Based confusion matrix

Table III. Gender-Based Classification Report

Gender	Precision	Recall	F1-Score	Support
female	0.94	0.98	0.96	148
male	0.98	0.94	0.96	140
accuracy			0.96	288
macro avg	0.96	0.96	0.96	288
weighted avg	0.96	0.96	0.96	288

C. *Emotion Recognition*

When the analysis was focused on emotion recognition, disregarding gender, the CNN achieved an accuracy of 63% and a weighted average F1-score of 0.63. The model performed well on emotions like "surprise", "calm", and "angry", achieving F1-scores of 0.76, 0.69, and 0.72, respectively. However, it struggled with "neutral" and "sad" emotions, which had lower F1-scores of 0.29 and 0.50, respectively. These results suggest that some emotions, as mentioned earlier, particularly those with subtle vocal variations, pose challenges for the model. The analysis highlights the complexity of emotion recognition tasks and the potential for improvement by incorporating additional features or refining the CNN architecture. The detailed confusion matrix and classification report can be seen in Fig. 5 and Table IV.

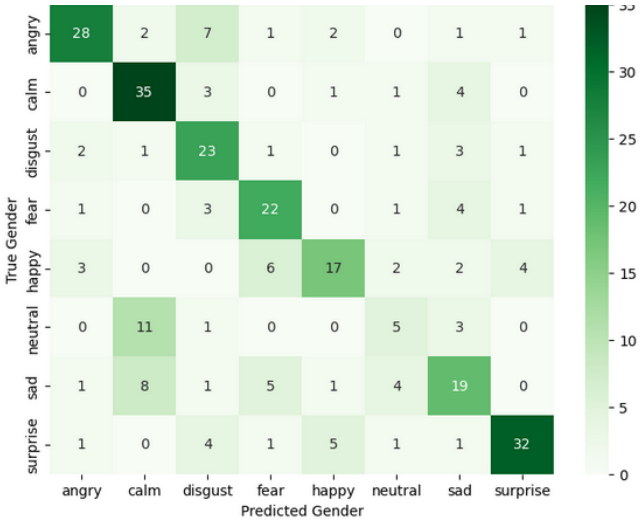

Fig. 5. Emotion-Based confusion matrix

Table IV. Emotion-based Classification Report

Emotion	Precision	Recall	F1-Score	Support
angry	0.78	0.67	0.72	42
calm	0.61	0.80	0.69	44
disgust	0.55	0.72	0.62	32
fear	0.61	0.69	0.65	32
happy	0.65	0.50	0.57	34
neutral	0.33	0.25	0.29	20
sad	0.51	0.49	0.50	39
surprise	0.82	0.71	0.76	45
accuracy			0.63	288
macro avg	0.61	0.60	0.60	288
weighted avg	0.63	0.63	0.63	288

IV. Conclusion

This project presented a speech emotion recognition system using the RAVDESS dataset, with 2D MFCC features extracted from audio files. Two models, SVM and CNN, were evaluated for their performance, and the results demonstrated that the CNN outperformed the SVM, and achieved higher accuracy and F1-scores across most emotion categories. While the CNN performed well in emotion recognition, its effectiveness varied across different emotions, with notable challenges in recognizing subtle categories like "neutral" and "sad".

To further enhance the performance of the system, several areas of improvement can be explored. Incorporating additional datasets with similar audio features could help in training more generalized models by having more audio files in more ranges. Moreover, extracting a broader range of features beyond MFCCs, such as spectrograms or pitch-related features, could provide richer input for the model. Finally, fine-tuning the CNN architecture, including experimenting with deeper layers or alternative configurations, may yield better results and improve the model's ability to capture complex patterns in speech emotion recognition tasks. These enhancements would contribute to a more robust and versatile emotion recognition system.

This project highlights the potential of speech emotion recognition in enhancing various applications. However, its deployment must be accompanied by careful consideration of

ethical issues. The collection and processing of audio data raise privacy concerns, particularly if users are unaware of their data being analyzed. Furthermore, the technology could be misused in surveillance or manipulative advertising, creating a need for clear guidelines and ethical standards. Future work should address these challenges by incorporating privacy-preserving mechanisms, to ensure transparency, and promoting fairness in model performance across different demographic groups.

V. References

[1] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)”, PLoS ONE, vol. 13, no. 5. Zenodo, p. e0196391, Apr. 05, 2018. doi: 10.5281/zenodo.1188976

[2] Wikipedia. Mel-frequency cepstrum. [Online]. Available: https://en.wikipedia.org/wiki/Mel-frequency_cepstrum [Accessed on Jan. 2, 2025]

[3] Devopedia, "Audio Feature Extraction," Version 8, May 23, 2021. [Online]. Available: https://devopedia.org/audio-feature-extraction. [Accessed: Jan. 2, 2025].