Speech Emotion Recognition: Clustering and Deep Learning Approach to Detect Conflicting Emotions through Vocal Expressions

Imtela Islam, Nuhash Kabir Neeha, Nuzhat Rahman
Supervisor: Dibyo Fabian Dofadar

BRAC UNIVERSITY
Inspiring Excellence

Abstract

As the dependence on automated systems increases, it is essential for machines to comprehend human emotions in order to enhance user experience and minimize failures in sectors such as virtual assistants, customer service, and emergency hotlines. Speech Emotion Recognition (SER) is crucial in this context, as it seeks to decode emotions expressed through vocal expressions. Nevertheless, current SER models face challenges in recognizing conflicting emotions when individuals express multiple feelings simultaneously. This research introduces a clustering and deep learning methodology that effectively identifies both primary and overlapping emotions, thereby making a more natural and emotionally intelligent human-computer interaction.

Dataset Description and Preprocessing

CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset
- 7,442 clips, 91 actors (48 male, 43 female)
- 6 emotions: Angry, Disgust, Fear, Happy, Neutral, Sad

RAVDESS: Ryerson Audio-Visual Database of Emotional Speech and Song
- 24 actors
- 7 emotions: calm, happy, sad, angry, fearful, disgust, and surprised
- Includes both speech and song

SAVEE: Surrey Audio-Visual Expressed Emotion
- 4 male British speakers
- 7 emotions: Angry, Disgust, Fear, Happiness, Sadness, Surprise, Neutral
- Detailed labeling of utterances
- Synchronized video/audio

TESS: Toronto Emotional Speech Set
- 2 female actors
- 200 target words
- 7 emotions: Angry, Disgust, Fear, Happiness, Pleasant Surprise, Sadness, Neutral

Preprocessing: Audio files were resampled, padded/truncated, and features such as MFCCs, Mel-spectrogram, and Chroma were extracted and standardized to fixed dimensions for model input.

Methodology

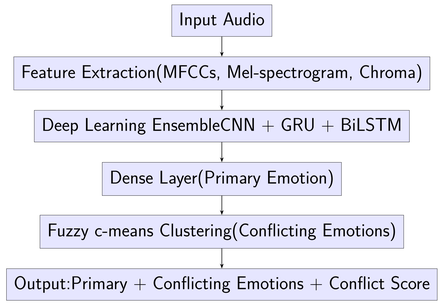

Figure 1. Flowchart of Methodology

Initial Model and Results

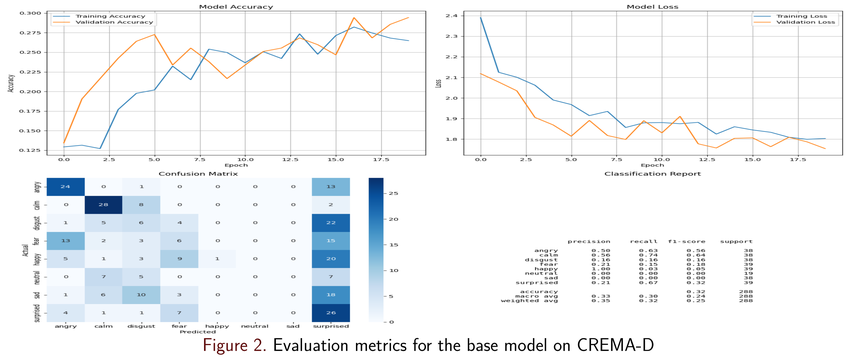

Figure 2. Evaluation metrics for the base model on CREMA-D

Model:
- Ensemble architecture combining CNN (for spectral features), GRU, and BiLSTM with attention (for sequential features)[1]
- Inputs: MFCCs, Mel-spectrogram, and Chroma features extracted from audio
- (here)Trained and evaluated on the CREMA-D dataset (7 emotions)

Results:
- Moderate accuracy (32%) on the test set
- Confusion matrix shows frequent misclassification among low-arousal emotions (e.g., sad, neutral, disgust)
- Training curves indicate stable learning with a minor generalization gap
- Model struggles to distinguish subtle differences between similar emotions, especially in challenging classes

Proposed Model and Results

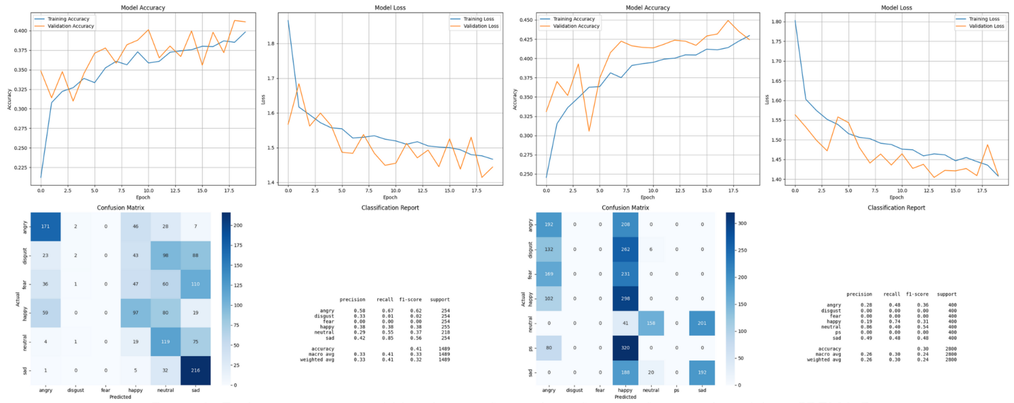

Figure 3. Evaluation metrics and Visualisation for applying fuzzy c after initial model on CREMA-D

Model:
- Added Fuzzy C-means clustering to classify emotions better
- Inputs: MFCCs, Mel-spectrogram, and Chroma features extracted from audio
- (here)Trained and evaluated on the CREMA-D dataset (7 emotions), in addition to cross corpus train-test using CREMA-d and TESS respectively

Results:
- Improved accuracy (41%) on the test set (CREMA-D), and 30% on the cross corpus run (trained on CREMA-D, tested on TESS)
- Confusion matrix shows significant improvement on the low-arousal emotions
- Training curves indicate more stable learning with a lesser generalization gap
- Model distinguishes subtle differences between similar emotions significantly better

Analysis

Final model – CREMA-D

Accuracy: Moderate (41%).

Confusion Matrix: Misclassification occurs among fear, disgust, and sadness as a result of overlapping acoustic characteristics.

Training Curves: Learning is stable, despite a minor generalization gap.

Analysis: The performance is moderate, due to the dataset's large size and diverse emotional intensity. It still faces difficulties in distinguishing subtle differences between low-arousal emotions (e.g., sadness versus neutral). This further emphasizes the challenges associated with nuanced emotion recognition, even within well-annotated datasets.

Final model cross corpus with CREMA-D

Accuracy: Notable decline (approximately 30%).

Confusion Matrix: Higher misclassification rates (e.g., distinguishing between happy and neutral).

Fuzzy Clustering: Indicates emotional overlap, particularly within TESS's older female vocal samples.

Analysis: Poor generalization attributed to demographic discrepancies (CREMA-D: varied ages and genders; TESS: exclusively older females). The distinct vocal characteristics of TESS contrast with the more naturalistic recordings found in CREMA-D. This highlights the cross-corpus difficulties faced in practical applications.

Conclusion

Our hybrid model enhances speech emotion recognition (SER) by identifying overlapping emotions; however, it encounters generalization gaps stemming from biases within the dataset.

Future research will focus on attempting to broaden demographic diversity, and enhancing interpretability for practical applications.

References

[1] E. Lieskovská et al., *A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism*, Electronics, vol. 10, no. 10, p. 1163, 2021. https://www.mdpi.com/2079-9292/10/10/1163

imtela.islam@g.bracu.ac.bd, nuhash.kabir.neeha@g.bracu.ac.bd, nuzhat.rahman@g.bracu.ac.bd