Fuzzy-Based Emotional Conflict Detection in Speech Using a Hybrid Deep Ensemble Framework

Nuhash Kabir Neeha
Department of Computer Science and Engineering
Brac University
Dhaka, Bangldesh
nuhash.kabir.neeha@g.bracu.ac.bd

Nuzhat Rahman
Department of Computer Science and Engineering
Brac University
Dhaka, Bangldesh
nuzhat.rahman@g.bracu.ac.bd

Imtela Islam
Department of Computer Science and Engineering
Brac University
Dhaka, Bangldesh
imtel.islam@g.bracu.ac.bd

***Abstract*—This paper presents a new hybrid deep learning framework aimed at identifying emotional conflicts in speech. This goes beyond standard Speech Emotion Recognition (SER) approaches by detecting overlapping or simultaneous emotions. The model integrates convolutional neural network (CNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) with Attention) along with Transformer (Wav2Vec2) elements within an ensemble framework, along with a Fuzzy C-Means (FCM) clustering module to evaluate emotional ambiguity. The framework was evaluated on enhanced versions of the RAVDESS, SAVEE, and CREMA-D datasets. The findings show that the ensemble achieved accuracies of up to 88.75%, with a conflict detection rate ranging from 15 to 21%, and an average fuzzy partition coefficient of 0.84, indicating clear emotion-conflict boundaries. These findings highlight that fuzzy conflict modeling provides a useful method for assessing emotional overlap, improving interpretability and reliability in affective computing.**

***Index Terms*—Speech Emotion Recognition (SER), hybrid deep learning, emotional conflict detection, fuzzy C-means (FCM), ensemble learning, affective computing**

I. Introduction

Human emotions are rarely one-dimensional, individuals frequently show mixed or opposing emotions like anger combined with disgust or sadness with happiness. Conventional SER systems categorize emotions separately, missing these overlaps. This research addresses that limitation by proposing a fuzzy-based emotional conflict detection framework, which detects both primary and secondary emotions and quantifies emotional ambiguity.

A. *Contributions:*

- A hybrid deep ensemble combining CNN, Transformer, and RNN-BiLSTM with GRU and attention mechanisms branches for robust emotion representation.
- Integration of fuzzy clustering to measure emotional overlap through conflict scoring.
- Empirical validation of emotional conflict detection across multiple datasets.
- Establishment of a new interpretability layer connecting categorical and continuous emotional models.

II. Related Work

Research on Speech Emotion Recognition (SER) includes both traditional signal-processing methods and contemporary deep learning techniques. Here, we categorize previous studies into four main areas and offer brief critical evaluations of key papers that directly inspire the current research.

A. *From handcrafted features to deep representations*

In the early days of speech emotion recognition (SER), researchers relied on manually created acoustic features like Mel-Frequency Cepstral Coefficients(MFCC), prosodic, and spectral characteristics, using traditional classifiers such as support vector machines (SVM) and Hidden Markov Model (HMM). These techniques set important benchmarks but struggled with generalizing across different datasets and had limited ability to model time variations [1] [2]. With the rise of deep learning, CNNs began to capture distinct spectral and spatial patterns from spectrograms, while recurrent networks like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) focused on modeling temporal changes. More recently, self-supervised methods and Transformer-based embeddings, such as Wav2Vec2, have offered rich contextual representations that enhance performance on standard datasets [3] [4]. However, despite these advancements, many deep learning methods still consider emotions as single, unchanging labels, which means they do not adequately capture mixed emotions [5].

B. *Feature fusion and ensemble strategies*

A significant body of research indicates that combining different feature types or model structures (like CNN+RNN or spectral + Self-Supervised Learning (SSL) embeddings) results in small yet steady improvements in robustness and performance across different datasets [4] [5]. Using ensemble averaging and majority voting helps decrease variability between runs and enhances stability, although this often leads to greater model complexity and higher inference costs. Notably, many fusion studies fail to incorporate explicit ambiguity diagnostics, which limits the understanding of mixed emotional states [4].

C. Conflicting or mixed-emotion detection representative studies

- **ConflictNet [1]:** Presents a complete method for estimating conflict intensity from audio and shows a strong correlation with human ratings. However, it fails to connect intensity estimates to distinct secondary emotions, which restricts its usefulness for systems that require understandable alternative labels.
- **Multiple models fusion for multi-label SER [4]:** Shows that multi-label architectures and fusion enhance Unweighted Average Recall (UAR) and identify co-occurrence patterns, but these approaches focus on multi-label outputs instead of graded membership values that indicate levels of overlap.
- **Cross-modal conflict estimators (visual/DCNN studies) [8]:** Demonstrate effective conflict estimators in facial and visual data, but differences in modalities limit their direct use in audio-only SER.
- **Mixed-EVC (mixed emotion voice conversion) [2]** and **speech synthesis with mixed emotions [3]:** Present solid proof of the perceptual reality of blended emotions via synthesis and conversion, but emphasize generation/control instead of detection or understandable classification in recognition processes.

These studies show that mixed emotions are real and can be measured, but previous detection methods either give a continuous intensity without categorizing it or concentrate on synthesis instead of clear recognition [1] [2] [3] [4].

D. Interpretability, uncertainty estimation, and fuzzy approaches

Interpretability and uncertainty estimation in SER have been tackled using confidence thresholds on softmax outputs, Bayesian methods, and specialized uncertainty estimators [9]. Fuzzy clustering, particularly through FCM, provides graded membership values near class boundaries and has been utilized in similar affective tasks to highlight ambiguity. However, there are few SER studies that merge multiview ensemble outputs with FCM to generate both strong categorical predictions and clear continuous indicators of overlap. Our research takes this combined approach, incorporating feature-specific branches, ensemble fusion, and FCM post-processing to produce categorical labels along with interpretable membership diagnostics [2] [9].

E. Summary of gaps motivating this work

The literature therefore highlights three converging gaps:

- most systems still assume single, static emotion labels
- fusion or ensemble strategies improve accuracy but rarely provide interpretable diagnostics for ambiguity
- mixed-emotion synthesis or intensity estimation exist largely separate from categorical detection methods

The suggested multiview ensemble with FCM process combines these elements by generating strong primary labels and revealing varying secondary memberships for unclear statements, tackling the performance and interpretability issues mentioned earlier [3] [4] [9].

III. Methodology

A. Framework Overview

The proposed Emotional Conflict Detection Framework consists of four modules (Fig. 1):

- **Data Augmentation:** Noise, pitch shifting, time stretching, SpecAugment, and volume scaling balance emotion classes.
- **Feature Extraction:** Extracts MFCCs, Mel-spectrograms, Chroma, spectral features, and Wav2Vec2 embeddings.
- **Additional features:** ZCR, spectral dependencies etc.
- **Hybrid Ensemble Model:**
  - **CNN branch:** Captures spectral–spatial features.
  - **RNN branch (BiLSTM–GRU with Attention):** Models temporal emotion dynamics.
  - **Transformer branch (Wav2Vec2):** Captures contextual dependencies.
  - **Fusion Layer:** Merges embeddings into a 512-dimension vector with residual dense layers and dropout regularization.
  - **Adversarial Speaker Layer (GRL):** Minimizes speaker bias for generalization.
- **Sequential training & ensemble selection:**
  - Three independent hybrid models (same architecture) were trained sequentially with differing hyperparameters (e.g. learning rates, batch sizes)
  - The independently trained runs produced varied individual test accuracies
  - Top performing checkpoints from these runs were combined where predictions were merged using ensemble averaging of probabilities and a complementary majority-voting scheme to produce stable, combined outputs
  - The combined probability outputs were then forwarded to the fuzzy post-processing stage for conflict analysis
- **Fuzzy Conflict Analysis:**
  - Employs Fuzzy C-Means clustering (m=2, n=3) on model probabilities.
  - Computes $ConflictScore = 1 - Confidence(PrimaryEmotion)$
  - Detects secondary/conflicting emotions when confidence < 0.5 and secondary probability > 0.3.

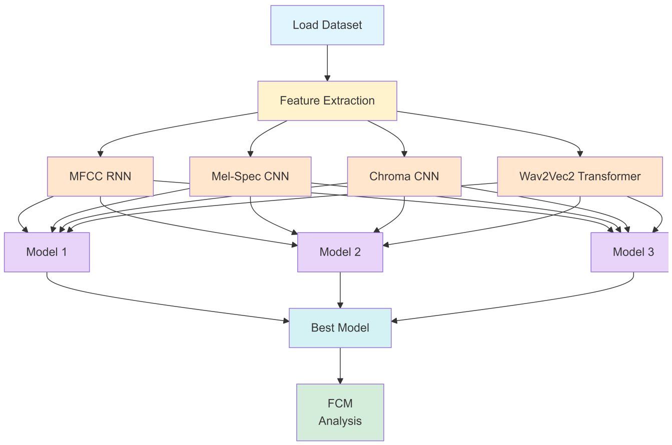

Fig. 1. Architectural Flowchart

IV. Experimental Setup

A. Datasets

- **RAVDESS:** 2,401 utterances
- **SAVEE:** 1,596 utterances
- **CREMA-D:** 16,801 utterances

Each dataset was augmented to balance classes and improve generalization.

B. Training Parameters

- Framework: TensorFlow, Librosa, Scikit-Fuzzy
- Optimizer: Adam (lr = 0.001)
- Epochs: 20-30, Batch size: 16-64
- Speaker loss weight (λ): 0.1
- Metrics: Accuracy, F1-score, Fuzzy Partition Coefficient (FPC), Cohen's Kappa (κ), and Conflict Rate (%)

V. Results and Analysis

A. Primary Emotion Classification

The proposed ensemble model showed robust performance for recognising primary emotions at over 80% accuracy across all the three augmented datasets used: with RAVDESS at 85.56%, CREMA-D at 85.71% and SAVEE at 88.75%. This, along with the average F1-score of 0.85 combined with the Cohen's Kappa averaging at 0.77, indicates that the ensemble model learns meaningful patterns instead of memorising, hence is reliable and has strong generalizability for all the nuances across corpora.

B. Emotional Conflict Detection

The moderately high fuzzy partition coefficient values ranging from 0.81 to 0.84 across the datasets suggests moderate fuzziness at class boundaries. While the conflict rates vary by corpus - with the highest being on the most diverse dataset, CREMA-D, and the lowest on the music based RAVDESS - the conflict scores are more comparable ranging from 0.44 to 0.48, which indicates that the model can successfully capture the existence of conflict in the corpora.

Additionally, when confidence was higher (above 0.8), the accuracy rose to over 96%, while at lower confidence levels it went down to 41%, matching the fuzzy conflict regions. This shows that the model successfully quantifies between certainty and uncertainty, once again proving that it is reliable, and has the ability to detect ambiguous emotional states.

TABLE I
Primary Emotion Classification

Dataset	Ensemble Accuracy%	Mean Individual Accuracy%	F1-Score	Cohen's κ
RAVDESS	85.56	81.69	0.85	0.78
SAVEE	88.75	80.69	0.86	0.79
CREMA-D	85.71	83.63	0.84	0.77

TABLE II
Emotional Conflict Detection

Dataset	Fuzzy Partition Coeff.	Conflict Rate%	Avg. Conflict Score	Most Conflicted Emotions
RAVDESS	0.84	15.28	0.46	Disgust, Neutral, Happiness
SAVEE	0.82	21.25	0.48	Anger, Fear, Neutral
CREMA-D	0.81	22.50	0.44	Sadness, Disgust, Neutral

- High-confidence samples (>0.8) $\rightarrow$ 96% accuracy
- Low-confidence (<0.5) $\rightarrow$ 41% accuracy (matches fuzzy conflict regions)

C. Comparative Performance

Compared with prior studies [1] [2] [4] the proposed framework achieved:

- +1–2% higher accuracy
- New metric: Conflict Rate (%)
- Comparable interpretability (FPC $\approx$ 0.84 vs PCC $\approx$ 0.85)

VI. Discussion

A. Insights

- Emotional conflicts align with known acoustic overlaps (e.g., *anger–disgust*)
- Fuzzy scores provide a continuous representation of mixed emotions, bridging categorical and dimensional emotion theories
- Ensemble synergy improved robustness by 4-6% over the best single model

B. Implications

- Enables explainable emotion AI through uncertainty quantification
- Applicable in therapy, sentiment-aware robotics, and empathetic chat systems

VII. Limitations and Future Work

A. Limitations

- Limited speaker diversity in datasets
- Synthetic augmentation may not fully mimic real-world variation
- Ensemble model increases computational cost

B. Future work

Future work will include:

- Real-time conflict-aware emotion detection
- Cross-cultural dataset expansion
- Multimodal integration (speech + facial cues)
- Human perception studies on fuzzy conflict validity

VIII. Conclusion

This study introduced a hybrid ensemble Speech Emotion Recognition (SER) model combining CNN, Transformer, and LSTM architectures with fuzzy conflict detection to address overlapping emotions in speech. Assessed using the augmented RAVDESS, SAVEE, and CREMA-D datasets, the system reached a high level of accuracy and better interpretability by combining fuzzy clustering with speaker-adversarial learning. This framework successfully measures emotional ambiguity, which boosts reliability in intricate vocal expressions. Future research could investigate multilingual, real-time, and multimodal extensions to broaden its use in emotionally aware human to machine interactions.

Acknowledgment

We would like to express our sincere gratitude to our supervisor, Mr. Dibyo Fabian Dofadar, our supervisor, for his invaluable guidance, patience, and continuous support throughout this research. His insights and encouragement have been essential to the completion of our work.

We also extend our appreciation to Dr. Md. Golam Rabiul Alam, Professor and Thesis Coordinator, and Dr. Sadia Hamid Kazi, Chairperson of the Department, for their academic supervision and for providing a supportive research environment.

Finally, we are deeply thankful to our peers, families, and friends for their encouragement, understanding, and moral support during this journey.

References

[1] Vandana Rajan, Alessio Brutti, and Andrea Cavallaro. Conflictnet: End-to-end learning for speech-based conflict intensity estimation. *IEEE Signal Processing Letters*, 26(11):1668–1672, 2019.

[2] Kun Zhou, Berrak Sisman, Carlos Busso, Bin Ma, and Haizhou Li. Mixed-evc: Mixed emotion synthesis and control in voice conversion. 2023.

[3] Kun Zhou, Berrak Sisman, Rajib Rana, Björn Schuller, and Haizhou Li. Speech synthesis with mixed emotions. 08 2022.

[4] Anwer Slimi, Nafaa Hafar, Mounir Zrigui, and Henri Nicolas. Multiple models fusion for multi-label classification in speech emotion recognition systems. *Procedia Computer Science*, 207:2875–2882, 2022. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022.

[5] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn Schuller. Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 14(2):1634–1654, 2023.

[6] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992.

[7] James A. Russell. A circumplex model of affect. *J. Pers. Soc. Psychol.*, 39(6):1161–1178, December 1980.

[8] Nikodem Rybak and Daniel J. Angus. Tracking conflict and emotions with a computational qualitative discourse analytic support approach. *PLOS ONE*, 16(5):1–29, 05 2021.

[9] Mohammad Mahdi Rezapour Mashhadi and Kofi Osei-Bonsu. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *PLOS ONE*, 18(11):1–13, 11 2023.