VCET - IIIC - DS & DL Internship - 2024-25

Speech Emotion Recognition for Distress Detection in Helplines

Proof-of-Concept Using MFCC, Chroma Features and SVM

Zubia Sarang[1] and Moin Khan[2]

[1]Department of AI & DS, VCET
[2]Department of EXTC, DBIT

July 6, 2025

Abstract

This study presents a proof-of-concept Speech Emotion Recognition (SER) model aimed at detecting distress-related emotions such as anger, fear, and sadness in speech signals. The proposed system is motivated by applications in child abuse helplines and support services for individuals in abusive relationships, where timely identification of distress is critical. We used the RAVDESS and CREMA-D datasets to construct a corpus of 4389 audio samples, which were preprocessed and augmented. MFCC and Chroma features were extracted to form 52-dimensional feature vectors, which were used to train a Support Vector Machine classifier. The model achieved a test accuracy of 75%, with particularly strong performance in detecting anger and sadness. These findings demonstrate the viability of using lightweight, interpretable SER models for real-time distress detection in social support contexts.

Keywords: Speech Emotion Recognition, Distress Detection, MFCC, SVM, Abuse Helplines

1 Introduction

Emotional distress in spoken communication is often a key indicator of abuse, trauma, or crisis. In helplines supporting children and individuals in abusive relationships, identifying distress early can lead to more effective interventions. Manual monitoring of calls is impractical at scale, necessitating automated methods to detect emotions from speech.

This paper proposes a Speech Emotion Recognition (SER) model focused on detecting *anger*, *fear*, and *sadness* — emotions commonly expressed by victims of abuse. By leveraging MFCC and Chroma features, and training a Support Vector Machine (SVM) classifier, we demonstrate that even a lightweight model achieves meaningful accuracy on standard emotional speech datasets.

2 Materials & Methods

We combined two publicly available emotional speech datasets: RAVDESS and CREMA-D, filtering for three target classes (*angry*, *fear*, *sad*). Data augmentation techniques, including Gaussian noise addition, time and pitch shifting, and stretching, were applied to increase variability.

For each sample, 40 MFCC coefficients and 12 Chroma features were extracted, resulting in a 52-dimensional feature vector. Features were standardized, and labels encoded.

The dataset was split into 80% training and 20% testing sets. A Support Vector Machine classifier with RBF kernel was trained on the training set and evaluated on the test set.

Spectrograms of representative samples from each emotion were also generated for qualitative analysis.

3 Results

The proposed model achieved an overall test accuracy of 75%. Class-wise F1-scores were as follows: *angry* — 86%, *fear* — 64%, *sad* — 75%. The confusion matrix (Figure 2) shows that fear was the most difficult class to predict accurately, whereas anger was detected with high precision and recall.

Figures 1 showcase spectrograms of representative audio samples, illustrating distinct patterns across emotions.

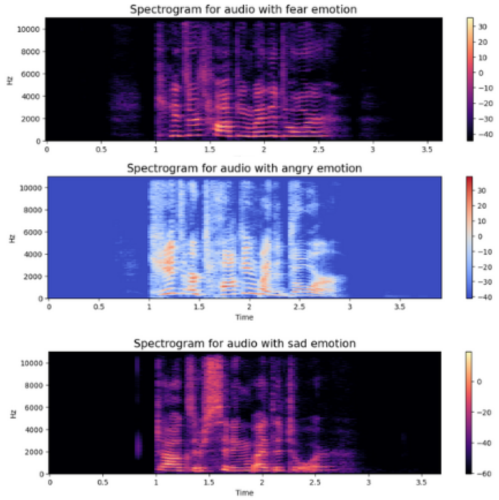


Figure 1: Spectrograms of representative samples: (a) Fear, (b) Angry, (c) Sad.

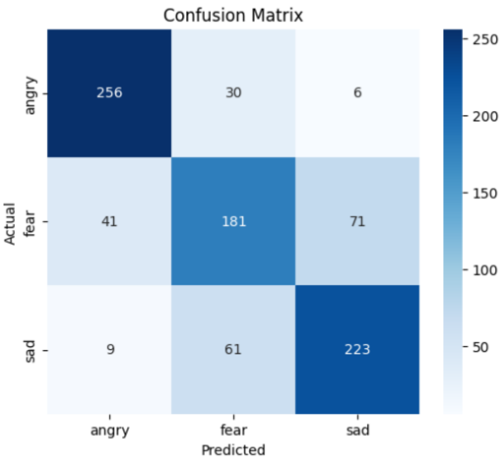
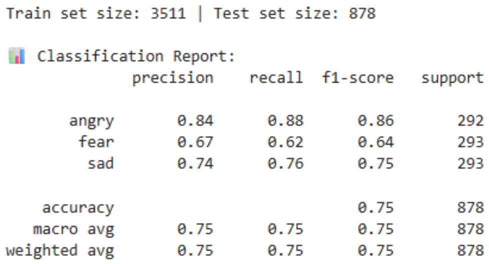


Figure 2: Confusion matrix for the SVM classifier on the test set.

4 Discussion

The results confirm the effectiveness of MFCC and Chroma features with SVM in detecting distress-related emotions. Anger detection was strongest, likely due to its more pronounced acoustic signature, while fear was less accurately detected, potentially due to overlapping features with sadness.

For deployment in real-world helplines, further improvements are needed to address fear detection, possibly by including prosodic or temporal features, or by employing deep learning approaches.



```
Train set size: 3511 | Test set size: 878

📊 Classification Report:
              precision    recall  f1-score   support

       angry       0.84      0.88      0.86       292
        fear       0.67      0.62      0.64       293
         sad       0.74      0.76      0.75       293

    accuracy                           0.75       878
   macro avg       0.75      0.75      0.75       878
weighted avg       0.75      0.75      0.75       878
```

Figure 3: Test set accuracy across classes.

Despite limitations, the current model demonstrates the feasibility of integrating SER systems into support services, enabling quicker response to victims of abuse.

Conclusions

This study establishes a baseline SER model for detecting distress in helpline contexts. Future work will explore deep neural networks and context-aware models to improve classification of subtle emotions like fear, and extend the system to real-time applications.

Acknowledgements

We thank the faculty of the Department of AI & DS, VCET, for their support and guidance throughout this project.

References

[1] Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13.5 (2018): e0196391.

[2] Cao, Hu, et al. "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset." IEEE Transactions on Affective Computing 5.4 (2014): 377-390.

[3] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.