International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Emotion Stream: A CNN-Based Speech Emotion Recognition System for Real-Time Music Recommendation

Mrs. Yasmeen Viqar [1], Kamran Ajaz Shah [2], Shahid Zahoor Koul [3], Zarnain [4]

[1] *Assistant Professor, Department of Computer Science and Engineering, SSM College of Engineering,193121*
[2,3,4] *Student, Department of Computer Science and Engineering, SSM College of Engineering, 193121*

Abstract - Emotions expressed in speech play a vital role in how people communicate and interact. Teaching machines to recognize these emotions can make digital systems more adaptive and engaging. This study introduces *Emotion Stream*, a web-based application that identifies the emotional state of a speaker and suggests suitable music or podcasts in real time. The system relies on Mel Frequency Cepstral Coefficients (MFCCs) to extract meaningful features from audio signals, while a Convolutional Neural Network (CNN) performs multi-class classification of emotions such as happiness, sadness, anger etc. A simple interface allows users to provide voice input, which is analyzed instantly, with results linked to Spotify's recommendation service for personalized content delivery. Testing indicates that CNN models capture the subtle variations in speech that reflect emotional tone, enabling reliable classification. By combining deep learning with real-time deployment, the application demonstrates how speech emotion recognition can be used to create emotion-aware recommendation systems. The work highlights the growing role of artificial intelligence in enhancing multimedia experiences and offers a framework that could be extended to other human–computer interaction domains.

Keywords: Speech Emotion Recognition, Convolutional Neural Network (CNN), Mel Frequency Cepstral Coefficients (MFCC), Emotion-Aware Recommendation, Spotify Integration, Human–Computer Interaction.

1.INTRODUCTION

Emotions shape the way people think, interact, and respond to the world around them. Speech, in particular, carries subtle cues about a person's emotional state, expressed through tone, pitch, and rhythm. Teaching machines to recognize these patterns is a growing research challenge with applications in mental health monitoring, customer service, intelligent assistants, and personalized entertainment. Unlike text, which conveys meaning explicitly, speech requires systems to capture hidden emotional signals that vary across speakers and contexts, making automatic recognition a complex task.

Speech Emotion Recognition (SER) aims to address this challenge by combining signal processing techniques with machine learning models that can detect and classify emotions. Recent advances in deep learning have made it possible to achieve higher accuracy by learning intricate features directly from audio data. Still, real-time deployment and user-centric integration remain underexplored.

In this work, we present *Emotion Stream*, a web-based system that identifies emotions from speech and delivers mood-aligned music or podcast recommendations using Spotify. By integrating Convolutional Neural Networks (CNNs) with Mel Frequency Cepstral Coefficients (MFCCs), the platform demonstrates how SER can move beyond theory to deliver practical, emotionally intelligent multimedia experiences.

2.RELATED WORK

Speech Emotion Recognition (SER) has been an area of growing interest because of its potential to improve applications ranging from human–computer interaction to healthcare and personalized multimedia. The earliest approaches relied mainly on handcrafted acoustic features such as pitch, intensity, and Mel-Frequency Cepstral Coefficients (MFCCs). These were often classified using methods like Support Vector Machines (SVMs) or Gaussian Mixture Models (GMMs). For example, one study combined MFCCs with an ensemble of Gaussian kernel SVMs and achieved 75.79% accuracy [1]. Although useful, such approaches were limited by their dependence on manual feature engineering and their difficulty in adapting across speakers and recording environments.

With the rise of deep learning, SER research moved toward automated feature learning. Convolutional Neural Networks (CNNs) were applied to spectrograms, treating speech as a visual signal and learning meaningful features directly from data [2]. Transfer learning approaches, using pre-trained architectures such as AlexNet and VGG, showed further improvements by adapting image-based models to audio tasks [3]. Hybrid architectures, such as CNN–Bidirectional Long Short-Term Memory (CNN-BLSTM) models [4], proved effective at capturing both spatial frequency features and temporal dynamics. Similarly, lightweight models such as one-dimensional CNNs and Recurrent Neural Networks (RNNs) [5] demonstrated promise for sequence modeling. Comprehensive surveys of deep representation learning for SER [6] reinforce the value of these approaches in advancing the field.

Research has also drawn on insights from psychology. Ekman's model of six basic emotions [7] provided the theoretical foundation for many SER datasets and classification frameworks. Databases such as the Berlin Emotional Speech Database (EMO-DB) [9] have become benchmarks for training and evaluation. At the same time, newer approaches combining deep neural networks with Extreme Learning Machines (ELMs) [8] highlighted the benefits of hybrid classifiers. More recently, transfer learning techniques [10] have gained attention for their ability to adapt SER models across domains with limited training data.

Alongside technical progress, SER has been applied in practical systems. Emotion-based music recommendation engines [6] demonstrated how real-time emotion detection could guide playlist generation, showing the potential of SER to enrich everyday user experiences. Taken together, these contributions illustrate the field's transition from handcrafted feature engineering to deep learning–driven, application-oriented solutions, providing the foundation for our work on *Emotion Stream*.

3.SYSTEM ANALYSIS AND DESIGN

The *Emotion Stream* platform was developed with the goal of recognizing emotions in real time and linking them to personalized music or podcast recommendations. Before building the system, a feasibility study was carried out to confirm that the design was both technically and socially meaningful.

- **Technical feasibility**: The system used widely adopted tools such as Python, TensorFlow for deep learning, Librosa for feature extraction, Flask for backend deployment, and Spotify's API/SDK for integration. These technologies provided the necessary computational support while remaining lightweight for deployment.
- **Economic feasibility**: Since most of the tools were open source and the APIs offered free access for research purposes, the project remained low-cost and scalable.
- **Operational feasibility**: The platform was built around an intuitive interface. Users only needed to log in with Spotify and provide speech input, while the system handled pre-processing, feature extraction, classification, and playlist generation.
- **Social feasibility**: By aligning music with emotional states, the system adds value beyond entertainment—it promotes mood regulation and mental well-being.

The design of the system included several modules: audio pre-processing, MFCC feature extraction, CNN-based emotion classification, recommendation mapping, and backend session management. These modules were integrated using an Agile life cycle model, which allowed iterative refinement of both the user interface and the deep learning model.

4.IMPLEMENTATION

The implementation stage translated the design into a working application. After authenticating with Spotify (Fig. 4.1), users could record speech or upload an audio file (Fig. 4.2). The audio signal was then cleaned, normalized, and transformed into MFCCs. These features were passed into the CNN model, which consisted of convolutional layers to extract local frequency information, pooling layers to reduce dimensionality, and fully connected layers for classification.

The model was trained on publicly available emotional speech datasets, ensuring coverage of common emotions such as happiness, sadness, anger, and neutrality. Performance was evaluated through training curves and a confusion matrix (Fig. 4.3). Emotions with distinct acoustic signatures, such as anger and happiness, were recognized with higher confidence, while sadness and neutrality showed some overlap. Additional visualizations, such as wave plots highlighted the acoustic differences across emotion categories.

Finally, the classified emotion was mapped to Spotify's Web API, which retrieved playlists or podcasts reflecting the user's emotional state. These recommendations could be played directly in the browser, providing a seamless integration of SER and multimedia content delivery.

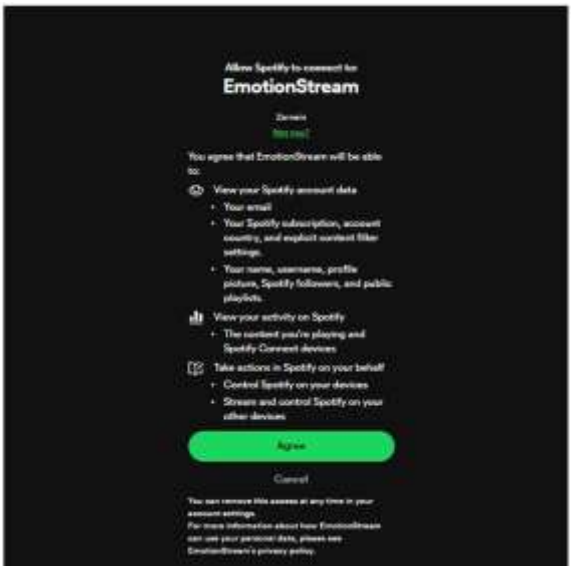

Fig 4.1: Spotify user authorization screen

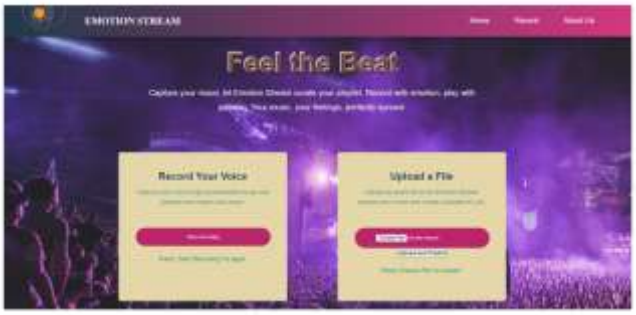

Fig 4.2: Record and upload audio interface

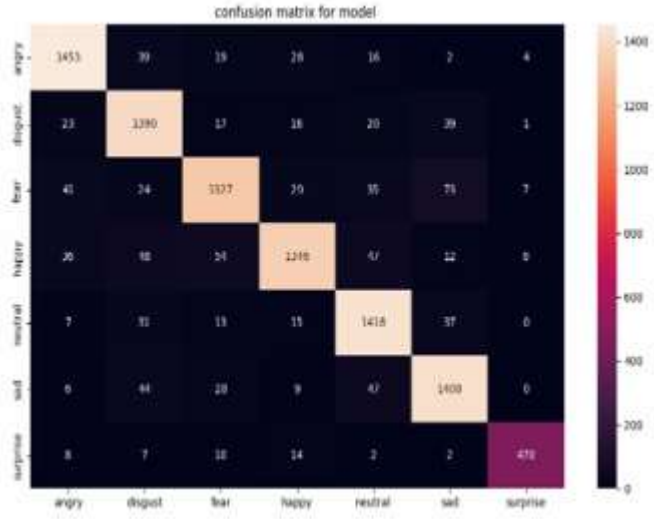

Fig 4.3 Confusion matrix of CNN model performance

5.RESULT AND DISCUSSION

The system successfully demonstrated its ability to detect emotions and provide contextually relevant recommendations. The result page (Fig. 5.1) displayed both the detected emotion and the suggested playlists or podcasts.

The CNN classifier proved effective in capturing patterns within speech. Analysis of the confusion matrix confirmed that

emotions like anger and happiness were classified with high reliability, while neutral and sad emotions occasionally overlapped due to their acoustic similarity. These outcomes align with earlier studies (Sec. 2), reinforcing the idea that CNNs combined with MFCCs are well suited for SER.

From the user's perspective, the Spotify integration added significant value. Instead of simply reporting the detected emotion, the system provided actionable recommendations. This ability to link recognition with personalized entertainment makes *Emotion Stream* more than a proof of concept—it is a functional tool for enhancing daily multimedia experiences.

Challenges remain, particularly around handling noisy environments and extending the emotional range. Expanding the dataset and including additional features such as pitch and prosody could improve performance further.

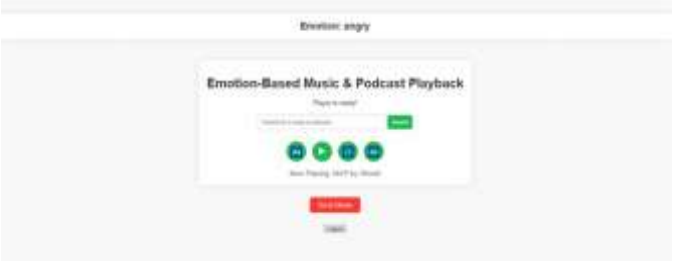

Fig 5.1 Result page displaying detected emotion and recommended playback

6. CONCLUSIONS AND FUTURE WORK

This paper introduced *Emotion Stream*, a speech emotion recognition platform that combines MFCC-based feature extraction with a CNN classifier and integrates with Spotify for personalized recommendations. The system achieved reliable classification accuracy and demonstrated how SER can be applied to create mood-aware multimedia experiences.

Future directions include broadening the range of recognized emotions to include categories such as surprise or disgust, and incorporating multimodal data (e.g., facial expressions or text sentiment) for greater robustness. Deploying the system on mobile platforms could make it more accessible, while advanced recommendation algorithms could enhance personalization.

Finally, beyond entertainment, this system has potential in areas such as mental health, where emotion tracking and adaptive interventions can provide real value.By bridging speech recognition with recommendation systems, *Emotion Stream* demonstrates how artificial intelligence can move beyond theoretical research to create emotionally intelligent technologies that enrich everyday life.

REFERENCES

[1] M. El Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[3] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B.W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2020.
[4] J. Lee and I. Tashev, "High-level feature representation using recurrent neural networks for speech emotion recognition," in *Proc. Interspeech*, pp. 1537–1540, 2015.
[5] S. Tripathi, S. Acharya, R. Sharma, and S.S. Mittal, "Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset," in *Proc. AAAI Conference on Artificial Intelligence*, 2017.
[6] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
[7] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.
[8] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, pp. 223–227, 2014.
[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, pp. 1517–1520, 2005.
[10] T. Zhang, W. Zheng, Z. Cui, and Y. Zong, "Deep transfer learning for speech emotion recognition," in *Proc. ICASSP*, pp. 5180–5184, 2018.