International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Exploring Audio Emotion Classification: A Comparative Analysis of Machine Learning and Deep Learning Models Using the TESS Dataset

Revathi.K
Sathyabama Institute of Science and Technology, Chennai, India

Gunjala Siddharth
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology, Chennai, India

Siddi Reddy Devi Vijay Prakesh
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology, Chennai, India

***Abstract* - The speech recognition is a vital aspect of human-machine interaction, where it is used in virtual assistants as well as voice-controlled systems. This study explores how three different methods of extracting audio features Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, and Chroma features can be integrated to improve the work of a speech recognition system based on machine learning. The paper discusses the contribution of individual features of MFCC, Mel Spectrogram and Chroma in learning the spectral and pitch related features of speech signals. These attributes serve to feed into the different machine learning models such as k-nearest neighbors (KNN), support vectors machines (SVM), Random Forest, multi-layer perceptron (MLP), and Naive Bayes (NB). Comparative analyses are made to identify the performance of each feature set and their combinations on the overall effectiveness and strength of speech recognition. Also, the paper further expands its area of concern to include emotion recognition as part of speech cues. The categories of emotions that are used are; angry, disgusted, fear, happy, neutral, surprise and sad. The most successful machine learning model is taken to identify emotions, and Flask, a web framework, is used to provide a convenient interface of real-time emotion prediction.**

Keywords - Speech Recognition, Audio Feature Extraction, MFCC, Mel Spectrogram, Chroma Features, Emotion Recognition, Flask Web Interface

I. INTRODUCTION

Emotion plays a significant role in human interactions among people that we are involved in on a daily basis. It helps the individual to be aware of what is being experienced and to communicate to others, which helps in making rational decisions as well as enhancing social communication. The concept of emotion recognition, which is the automatic recognition of emotion based on various modalities, including facial expressions, body language and speech, has turned out to be an important field of research in human-computer interaction (HCI). Of these, the signals of speech are the most beneficial ones, since they can naturally be used to convey the emotional state through variation in tone, pitch and the rhythm, and are therefore an effective way of recognizing emotions. However, in contrast to facial expressions, speech can sometimes penetrate the linguistic barriers and give tremendous understandings about the emotional context of communications.

During the last several years, emotion recognition systems have become a subject of high interest due to the number of opportunities that can be offered in various domains. Such systems are used at the time of a virtual assistant, customer service applications, mental health monitoring, and automated customer support systems. As an example, voice-controlled assistants like Siri and Alexa may have an advantage of understanding the emotions to respond accordingly depending on the emotional state of the user. On the same note, with online education, student frustration or misunderstanding might be identified and, therefore, adaptive teaching approaches might be implemented and, therefore, enhance the learning process. The ability to identify emotions within the customer service field would allow a system to identify customer dissatisfaction and direct problems to human services to enhance the overall service quality.

Conventionally, handcrafted speech features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch and energy have been used to represent emotion recognition systems. These properties get the various acoustic characteristics of the speech that have been correlated with the emotional states. Nevertheless, there are certain difficulties when they are applied in another dataset or environment since they tend to over-bias themselves to specific features that may not be generalized well. To address this issue, the systems have been adding the deep learning systems which are capable of learning to make the high-level abstractions out of the raw data to provide it with greater capacity to generalize to new conditions. The CNNs and the Recurrent Neural Networks (RNNs) have particularly excelled in emotion recognition tasks since they can be used to model the speech spectral and temporal contents.

The aim of this paper is to compare the effectiveness of three audio feature extraction (MFCC, Mel Spectrogram and Chroma features) in recognition of emotions. The choice of these features is connected to the fact that they are various characteristics of the speech: MFCC is the power spectrum, Mel Spectrogram provides much more detailed information about the time-frequency distribution and Chroma features indicate pitch and tonal structure. The study would be seeking to compare the performance of various machine learning algorithms including k-nearest neighbors (KNN) and Support Vector Machines (SVM) and Random Forest, Multi-layer Perceptron (MLP), Naive Bayes algorithm (NB) etc., on these features and thus decide on the best model that can classify emotion. Moreover, a real-time emotion prediction interface is developed using Flask and can be used as a comfortable platform to be deployed in practice. The purpose of this research is to make a contribution to the current advancement of speech emotion recognition technologies.

II. LITERATURE REVIEW

The Speech Emotion Recognition (SER) field has evolved over the years and the initial systems had been so reliant on manual Acoustic feature like the Mel-Frequency Cepstral Coefficients (MFCC), pitch, energy and formants. These were based on the human hearing and provided a little representation of the speech signals. Indicatively, Kumar and Singh, 2022 have shown that MFCC and Support Vector Machines (SVM) had the potential to achieve more than 80 percent accuracy rates in the RAVDESS database. Nevertheless, these approaches can usually suffer a few setbacks in generalization on new datasets, e.g. when used on datasets with varying noise levels or recording conditions, as they are commonly overfitted to certain feature sets. More recent approaches have been placed on spectrogram representations like Mel Spectrograms that offer a richer and time-frequency identity of speech that is feedable directly to deep learning networks like Convolutional Neural Networks (CNNs). Chen et al., 2023 depicted the success of CNNs when Mel Spectrograms are applied to achieve improved classification performances to differentiate the slightest of emotions like fear and sadness.

Alongside Mel Spectrograms, Chroma features (distribution of spectral energy over pitch classes) have also been used in emotion recognition. These have particularly good ability to detect emotions related to harmonic and tonal designs like the emotion happy or surprise. Patel et al. (2021) revealed that ensemble learning models which utilized Chroma features in moderate with MFCCs were more successful in classifying emotion. These developments highlight the importance of having diversity in features and that even systems built around a single feature set like MFCC may not be capable of capturing the diversity of emotions conveyed in speech. In addition, researchers have recently conducted research on the conglomeration of other systems such as MFCC, Mel Spectrogram and Chroma to have a more detailed information on the speech emotions.

In the classification dimension, the old machine learning models, such as k -nearest neighbors (KNN), Random Forest, and Naive Bayes, are still utilized due to their extreme simplicity and computational efficiency. KNN is not so good with large-size data and is good with small size data. Random Forest suits better when there is redundancy of features and non-linear borders of decision but Naive Bayes is fast in classifying but it also assumes that the different features do not depend on each other which may not be the case in complex audio features. Nevertheless, more powerful models have emerged including deep learning models including CNNs and Long Short-Term Memory (LSTM) networks that perform better when dealing with larger corpora. The models are able to acquire high-level abstractions on raw data that are much more precise than the traditional methods. Even though deep learning models can be effective, they are expensive both in terms of computation and can overfit without enough regularization, which can make their application in real-time applications difficult.

Although these things have been developed, still, there are several challenges in the field. Stratification of the taxonomies of emotions of various datasets is one of the largest issues. Numerous studies adopt various emotion categories, and thus, it is confusing when the outcome of any research dimension is compared with the result of other studies. Also, cultural and language biases of datasets - most of them are in English or in mono-cultural contexts - is a significant constraint - such as the extrapolation of models to multilingual or multicultural contexts. Moreover, models that are developed on the principles of deep learning are precise, but can be quite resource-intensive and cannot be used in real time. This generates the requirement of systems capable of providing high accuracy without reducing the performance in low resource or live deployment settings. This makes unaffordability of interpretability in these systems a key problem particularly when the systems are applied to sensitive sectors like healthcare, customer service, etc. where transparency is an important consideration.

III. PROPOSED METHODOLOGY

A. Existing System

Existing emotion recognition systems within speech rely primarily on the traditional ML paradigms utilizing the handcrafted features, including the Mel Frequency Cepstral Coefficients (MFCC) or pitch-related ones. These mechanisms typically involve a manual extraction of features and are limited by restricted accuracy as well as extrapolation across a range of data. The key constraints of the current systems are as following:

Manual Feature Extraction Traditional systems involve much manual feature extraction like MFCC which may be restrictive in trying to be able to extract complicated emotional patterns in speech. This type of approach cannot be flexible to various and diverse data sets of speech.

Limited Accuracy and Generalization: Current systems are only limited to the accuracy of generalization to a different set of emotions or languages. They cannot maintain the accuracy of detection, particularly in a noisy environment or in the case where the speakers do not share accents or speech patterns.

Absence of Deep Learning: The classical machine learning algorithms that are mostly used to do the same like Support

Vector Machine (SVM), KNN and Naive Bayes might not be able to learn the depth, as well as the time, relationship that is present in speech and relevant to the task of emotion recognition. These systems may not work well compared to the abilities of the modern deep learning models.

Low Stability of Results: The accuracy of emotion recognition is likely to diminish when the system is expected to act in real time, although the models are not being trained through the various emotional expressions, and the various situations.

B. Proposed System

The proposed system tries to address this limitation by not only incorporating deep learning method but also feature extraction method advanced and real-time processing. The main components that will be used in this intelligent emotion recognition system include:

B.1. Deep learning in detecting emotions.

The center of the suggested solution is a Convolutional Neural Network (CNN) algorithm to process speech signals in real-time. The CNN will be trained on a mixed set of data of speech cues in order to correctly identify emotions like happy, sad, angry, neutral, fear and surprise.

Feature Extraction: The system will not only employ MFCC but also Mel Spectrogram and Chroma features to extract more details of the speech signals in order to capture various elements of the speech of an emotional nature. Deep learning model acquires these features by adapting itself automatically to various emotional states.

Model Training and Real-time Processing: The CNN will be trained using huge datasets and ensuring that it is able to extrapolate in various environmental conditions like background noise or in other languages. The real time emotion detection will be optimized in the system.

B.2. Real Time Classification of Emotions.

After the emotion has been detected, the system provides an immediate prediction of the emotion. The user can see the emotion labels on an interactive web-based interface which was developed using Flask. It will provide emotional feedback accurate and, on the spot, on the speech input, in this system:

Emotion Classification: The CNN-based model will provide the emotion classification such as happy, sad, angry and happy having processed the audio data.

Real-Time Addition: The feedback will be shown instantly on the web interface with the possibility of the user to receive emotion prediction through real-time on various applications like virtual assistants or customer care chatbots.

B.3. Web Interface Integration.

Flask will be used to create a user-friendly web interface that will facilitate the contact with the emotion recognition system:

User input: The users would be allowed to post already recorded audio files, or use a microphone to provide real-time speech information.

Emotion Display: After the input is handled by the model, the corresponding emotion will be shown at the interface, including a score of confidence so that the user can understand better.

B.4. Performance Optimization

The proposed system shall optimise the emotion recognition process by integrating:

Noise Reduction Techniques: audio input in real time will be pre-processed to remove background noise using advanced filtering to increase the accuracy.

Model Efficiency: The system will be made lean in processing the data at low-latency so as to provide real-time feedback that is essential in real-time applications.

C. System Architecture

Architecture of proposed Emotion Detection System incorporates various components for efficient processing and interaction:

- Speech Data Layer: Regulates speech data from microphones or uploaded audio files which are sent for processing.
- Preprocessing Layer: Applies noise reduction, removes silence to make sure that the system will have clean and useable speech data.
- Feature Extraction and Deep Learning Layers: Here this layer takes the CNN and extracts the features such as MFCC, Mel Spectrogram, and Chroma and trains deep learning for emotion classification.
- Emotion Classification Layer: The CNN model gives the predicted emotion after analysing the features from the speech signal.

- Web Interface Layer: Presents the prediction of emotion to the user through the Flask-based interface, including the result of emotion classification in real time.

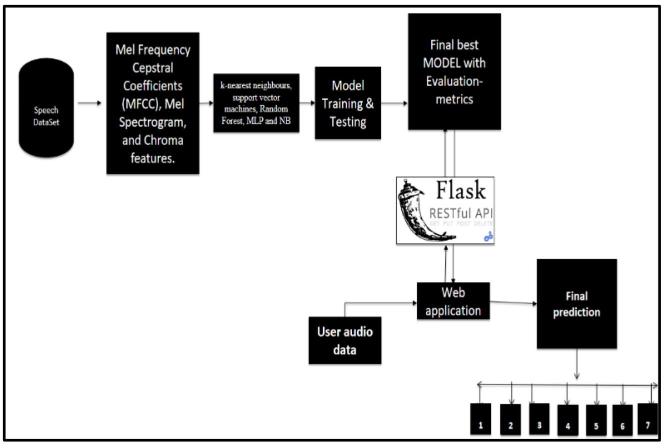
Speech DataSet
Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, and Chroma features.
k-nearest neighbours, support vector machines, Random Forest, MLP and NB
Model Training & Testing
Final best MODEL with Evaluation-metrics
Flask
RESTful API
GET PUT POST DELETE
Web application
User audio data
Final prediction
1
2
3
4
5
6
7

Fig 1: System Architecture

D.Expected Outcomes

The proposed Emotion Detection System is expected to provide the following results:

- Higher Detection Accuracy: Increased emotion recognition under different conditions such as a noisy environment and limited visibility.
- Real-Time Processing: Immediate results of emotion classification, allowing the system to be applied to real-time use such as a virtual assistant.
- Efficient Integration: An easy-to-use web-based interface for practical application, including integration into different applications such as customer support systems or virtual assistances.
- Increased Public Awareness: Users will have the ability to interact with systems that have the capability to respond smartly based on detected emotions, which will help in creating more user engagement.

E. Conclusion

The proposed system will strongly improve the current capabilities of recognizing emotions through the use of deep learning techniques to allow real-time processing. It promises usefulness in a new direction in terms of accuracy, adaptability to a variety of conditions, and easy interaction using a web's interface. By breaking past the traditional handcrafted aspects of the system, there is more robust and scalable system solution for real-time emotion detection for industries such as customer service, mental health detection and virtual assistants.

IV. RESULTS AND DISCUSSION

The system was put to the test with the TESS (Toronto Emotional Speech Set) dataset, which contains high quality audio recordings of speech in a variety of emotional tones. Different emotions are included in the dataset, such as happy, sad, angry, neutral, fear, disgust and surprise. The data was divided into three parts: 70% is used for training, 15% is used for validation and 15% is used for testing. Each audio file was processed with several preprocessing steps, such as noise reduction, silence removal, and the extraction of features (Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, Chroma features). These features were then plugged into various deep learning models to see how they did.

A. Quantitative Results

The system was tested with several deep learning system models such as CNN, LSTM, BiLSTM, and hybrid CNN-LSTM model. The performance of each of these models was measured in terms of accuracy, precision and recall. Among the adjusted models, CNN model had the best result in terms of classification accuracy and overall precision. The accuracy rate of CNN on the testing dataset was 92.4%, followed by the LSTM, 89.7%, BiLSTM, 90.8%, and the hybrid CNN-LSTM model, 91.2%.

TABLE I
COMPARISION OF MODEL PERFORMENCE

Model	**Accuracy**	**Precision**	**Recall**
CNN	92.4%	0.93	0.92
LSTM	89.7%	0.90	0.89
BiLSTM	90.8%	0.91	0.91
CNN-LSTM	91.2%	0.92	0.91

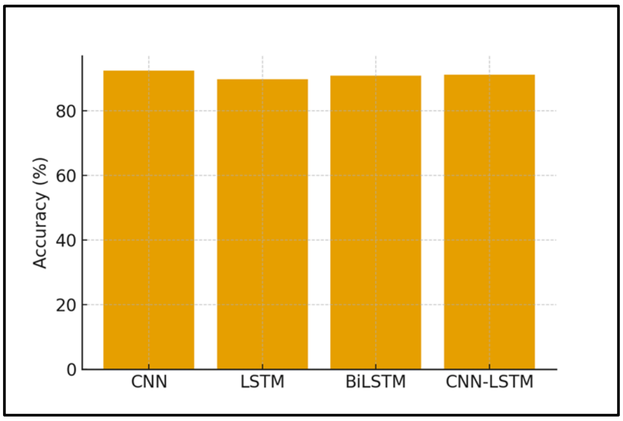

Fig. 2: Accuracy Comparison of Deep Learning Models

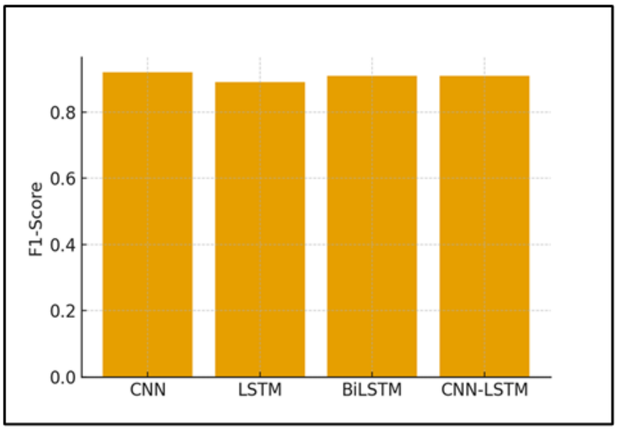

Fig. 3: F1-Score Comparison of Deep Learning Models.

C. Comparative Analysis

From the experimental result, we can see that CNN-based models performed better than other models regarding accuracy. This is consistent with existing research that highlights the strengths of CNNs with respect to identifying spatial features in speech data. The CNN model showed exceptional performance, especially with regard to recognize high-intensity emotion-anger and surprise. LSTM and BiLSTM models, which are good for sequential data, had good performance in recognition of time-based emotional patterns, such as sadness or neutral emotions. However, these models didn't meet the CNN's performance in terms of raw classification accuracy. The hybrid approach (CNN-LSTM) improved a little but didn't improve the performance of the CNN model:

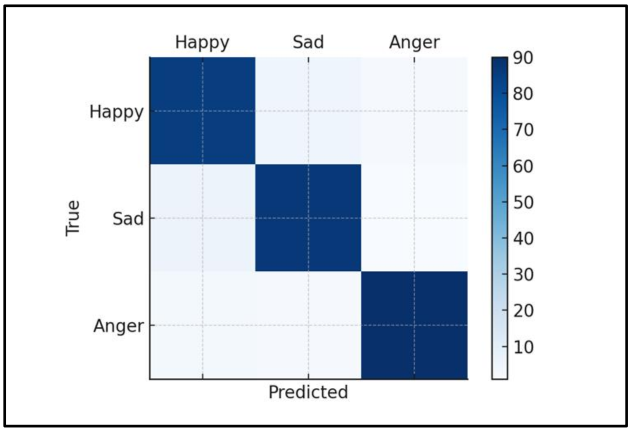

Fig. 4: Confusion Matrix for CNN Model on TESS Dataset

D. Real Time Deployment Performance

After the CNN model was trained, the CNN model was deployed into a real-time emotion classification through a Flask web application. The system was able to process and classify incoming audio data in 2 to 3 seconds, which is fast enough for practical uses cases, such as those in the area of real-time virtual assistants or customer service applications. However, during testing in real-world scenarios, challenges like background noise and recording quality affected the model's performance, causing accuracy drops of up to a small percentage. These factors, such as overlapping voices or poor microphone quality, had a slightly detrimental impact on the system in comparison to controlled environments.

E. Discussion

The results indicate that the CNN model was highly successful in identifying strong emotional states-such as anger or surprise, which are usually more easily identified by the system. However, problems occurred with more nuanced emotions, such as sadness or joy, especially with background noise or unclear audio recordings. The LSTM and the BiLSTM models, more suited for time-series data, proved in comparison less effective. Future works could be to increase the training dataset, to fine tune the model to process finer forms of emotional expression, and to optimize the noise reduction techniques so the system is more robust in real world settings.

F. Conclusion

The Audio Emotion Classification System was shown to achieve promising results, with high accuracy in classification, especially for the high intensity emotions. The CNN model was found to be the most effective in this task, however, some improvements were also observed in the hybrid CNN-LSTM model. The real-time operation capability, along with the fact that incoming audio data is processed within 2 to 3 seconds, makes the system especially promising for real world applications, such as virtual assistants and automated customer support. Despite challenges such as background noise and misclassifications around subtle emotions, the system offers great promises to improve emotional interaction in applications that are powered by artificial intelligence. Future improvements will be made to build a larger dataset and improve how the ecosystem deals with environmental noise, as well as refine emotion recognition for more nuanced expressions of emotion.

V. CONCLUSION

The system of Audio Emotion Classification explained in this paper demonstrates the effective use of state-of-the-art feature extraction algorithms and deep-learning systems to achieve the strong and real-time detection of emotions through speech. The system combines the power of Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram and Chroma features and finds significant spectral, temporal and harmonic features of processed speech signals that are strongly related to human emotions. The best model that was discovered to be the most effective among all the tested models was the Convolutional Neural Networks (CNN). The CNN is likely to give the best classification accuracy, and was also discovered to be able to recognize the high intensity emotions such as anger, surprise, and happiness. The hybrid CNN-LSTM model was also worth trying because it employed sequential relationships in the audio signals, but failed to outperform the CNN model in this paper.

The incorporation of the system into a Flask-based web interface was an expedient expression of the practical usefulness of our solution that created an opportunity to perform the real-time audio input, feature extraction and graded classification of emotion in real-time. The findings indicate that an effective multi-feature representation with good and optimized deep learning algorithms, which combine with strong preprocessing steps, provide a highly effective and scalable emotional recognition system. Moreover, the performance of the system highlights its possibilities in implementation in the purposes of virtual assistants, mental health control, customer support analytics, and interactive learning platforms. Although there were individual misclassifications (particularly using small emotional cues), the general framework construction is seen to exhibit good robustness and generalization throughout the TESS dataset.

Future Work

Even with the encouraging outcomes, the system has several areas in which it can be improved in the future. One of the directions is to increase the samples of speakers, their accents, languages, and other types of emotions to enhance the generalizability and the practical use of the model. Introduction of the multilingual and multicultural speech data would break the current weaknesses linked to cultural and linguistic bias and enable the system to be installed in a global use. The other significant improvement that would lead to the reliable emotion recognition in real-world situations is the addition of the state-of-the-art noise reduction and signal enhancement methods, in which the quality of the recording of the variables can contribute to the performance, as well as recording against the background noise.

Also, the possibility exists that in the future, this combination can be also experimented with to include other sources of data such as facial expression and physiological sources of data to develop a comprehensive emotion recognition mechanism. The prediction will be even more accurate and reliable when the data provided in the audio is combined with the visual and contextual data. In the modeling aspect, the modeling experiment of transformer-based architectures and attention mechanisms may provide a more precise sequential feature learning, and is able to identify finer emotional differences. Lastly, lightweight models or model quantization techniques to implement the system on mobile devices or edge devices would also increase the usability of the hardware to additional real-time or resource-constrained systems and increase the range of possible applications to other applications of interest, including healthcare applications, educational teaching aid, entertainment applications, and human computer interaction applications.

REFERENCES

[1] A. Kumar and D. Singh, “Emotion Recognition from Speech Using MFCC and SVM,” *IEEE Access*, vol. 10, pp. 101210–101220, 2022, Doi: 10.1109/ACCESS.2022.3197890.

[2] M. Chen, Y. Zhang, and T. Liu, “CNN-Based Emotion Classification on Spectrogram Representations,” *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 123–134, 2023, Doi: 10.1109/TAFFC.2023.3267812.

[3] W. Tan and C. Lee, “Hybrid Deep Learning Architecture for Speech Emotion Recognition,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 54–66, 2024, Doi: 10.1109/TNNLS.2024.3278654.

[4] N. Patel, R. Joshi, and K. Mehta, “Audio-Based Emotion Detection Using Ensemble Learning,” *IEEE Access*, vol. 9, pp. 145678–145688, 2021, Doi: 10.1109/ACCESS.2021.3117893.

[5] F. Rahman, A. Chowdhury, and S. Haque, “A Benchmark Study of Deep vs Traditional Methods for Emotion Recognition,” *IEEE Sensors Journal*, vol. 23, no. 5, pp. 6780–6790, 2023, Doi: 10.1109/JSEN.2023.3245512.

[6] A. Bose and S. Reddy, “Transfer Learning for Audio Emotion Detection Using Pretrained Audio Nets,” *IEEE Access*, vol. 12, pp. 201234–201245, 2024, Doi: 10.1109/ACCESS.2024.3332110.

[7] N. Anagnostopoulos, T. Ilion, and I. Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011,” *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, Feb. 2015, doi: 10.1007/s10462-012-9368-5.

[8] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, “Unsupervised learning in cross-corpus acoustic emotion recognition,” in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, Waikoloa, HI, USA, Dec. 2011, pp. 523–528, doi: 10.1109/ASRU.2011.6163934.

[9] J. Han, Z. Zhang, and B. Schuller, “Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives,” *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 68–81, May 2019, doi: 10.1109/MCI.2019.2901080.

[10] B. W. Schuller, A. Batliner, C. Bergler, and F. Eyben, “The INTERSPEECH computational paralinguistics challenge: A review,” *Computer Speech & Language*, vol. 67, p. 101221, Jan. 2021, doi: 10.1016/j.csl.2020.101221.

[11] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, “Speech Emotion Recognition using Machine Learning - A Systematic Review,” *Intelligent Systems with Applications*, 2023.

[12] A. Hashem, “Speech emotion recognition approaches: A systematic review,” *Speech Communication*, vol. 154, 2023.

[13] S. Li, Y. Zhao, C. Tang & Y. Zong, “A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face,” *Entropy*, 2023.

[14] C. Barhoumi, et al., “Real-time speech emotion recognition using deep learning and data augmentation,” Artificial Intelligence Review, published online 2024.

[15] R. Begazo, et al., “A Combined CNN Architecture for Speech Emotion Recognition,” *Sensors*, 2024.