International Journal of Computational Science and Engineering Research
ISSN: 3107 -8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol. 2, Issue. 2 (April - June), 2025 , 2025 , Pages: 30 - 34
Received: 24 February 2025 ; *Accepted:* 27 April 2025 ; *Published:* 28 June 2025
https://doi.org/10.63328/IJCSER-V2RI2P6
Original Paper : Applied Research – Compare and Contrast Research Paper

Survey on Emotion Detection via Audio Processing and Deep Learning

Boggadi Nagarjuna Reddy

Department of ECE, Viswam Engineering College, Madanapalle, Andhra Pradesh-517325, India;

** Corresponding Author: Boggadi Nagarjuna Reddy: nagarjuna@gmail.com*

Abstract: The expanding incorporation of speech and audio data in numerous applications is heavily dependent on Machine Learning (ML) and Artificial Intelligence (AI). The complexity of audio data analysis in machine learning stems from noise and variability, necessitating prepossessing and feature extraction techniques. These features enable machine learning algorithms to identify and recognize audio patterns. This study provides an overview of audio data processing and deep learning techniques used for emotion detection. This research presents a machine learning model for emotion detection, describes the model in detail, and explores anticipated results and potential avenues for further investigation.

Keywords: Audio data processing, Deep learning, emotion classification, machine learning, speech recognition.

1. Introduction

In the digital era, a vast quantity of data is being generated every second. The data comes from a variety of sources such as social media, educational institutions, entertainment firms, and many other domains. The data generated has the potential to contain valuable and relevant information that could be advantageous in a wide range of practical applications. These applications could have a positive impact on society and improve the overall well-being of individuals. Analyzing audio and video data from social media posts related to a natural disaster could be useful in preventing human casualties. Additional analysis of the audio and video data uncovers further potential options. A wide range of methods and techniques for analyzing audio and video data have been developed and published in recent research, primarily due to the work of numerous researchers and engineers. These methods are application-focused techniques that demonstrate how machine learning and deep learning techniques can be utilized to extract valuable insights from audio and video data. It was found during the investigation that the majority of the models utilize multi-word feature fusion methods, which incorporate audio-visual data to train deep learning models. This method of classification comes at a considerable expense. Applying audio signal analysis methods can be advantageous for optimizing resource usage. The project aims to investigate the techniques used in audio data processing and assess their effectiveness in terms of resource utilization. Additionally, examine the use of deep learning techniques in examining sound wave patterns. Enhanced deep learning methods have developed from artificial neural networks (ANN), providing multiple architectures that enable efficient and precise data analysis. In recent years, deep learning has proven its importance and practicality in a multitude of real-world uses. These methods are capable of dealing with intricate, substantial, and varied data analysis difficulties. These techniques can be utilized in medical care, image processing, EEG signal analysis, prediction, and other fields. The main emphasis of this research lies in the application of deep learning to audio signal analysis. Audio singles contain time-specific information, which is effectively modeled and extracted using time-dependent data approaches. This is achieved through the application of recurrent neural networks and their variants, including Long Short-Term Memory (LSTM) and bidirectional LSTM (bi-LSTM). The main goal of audio signal analysis is to isolate and extract the key features from speech data that are associated with emotional expression. This paper's scope and content have been summarized in the following section. The following section outlines the latest advancements in the field of audio data analysis through the application of deep learning methodologies. The review's conclusion is summarized using a table and its accompanying explanation. A proposed model is then explained and outlined for future design and development purposes. The final conclusion has been presented, and potential future research avenues have been outlined.

2. Literature Review

The proposed project aims to investigate the role of deep learning in the processing of audio data. This section presents a compilation of recent articles that have been reviewed to gain insight into the functioning of audio data processing. Relevant articles from established journals and conferences have been sourced via Google Scholar, with the most notable examples being highlighted in this section. The abbreviations mentioned in the literature surveyed are also compiled and presented in Table 1.

Table.1 List of abbreviations

ANN	Artificial Neural Networks
AER	Audio Emotion Recognition
AMFBP	Adaptive and Multi-level FBP
CNN	convolutional neural network
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
EEG	Electroencephalography
EMG	Electromyography
FBP	Factorized Bilinear Pooling
FCN	Fully Connected Network
G-FBP	Global FBP
HCI	Human-Computer Interaction
KNN	k-nearest neighbors
LSTM	Long short-term memory
MFCC	Mel-frequency cepstral coefficient
ML	Machine learning
MLP	Multilayer perceptron
PCA	Principal Component Analysis
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
SER	Speech Emotion Recognition
SVM	Support Vector Machines

A crucial aspect of human-computer interaction is taking into account the emotional state of the end user. This technology supports basic survival needs and has numerous practical applications in educational and medical settings. Techniques differ in their reliance on emotions, facial expressions, physiological responses, and neurological imaging. Text messaging service. A study by Abdullah et al [1] investigates emotion recognition signals using deep learning methods and reviews their relative uses. Research is being carried out on both multimodal and uni-modal approaches to enhance classification precision. The reliability of the results is contingent upon the number of emotional categories, characteristics, the classification method, and the data set employed. Improving research efforts would lead to a deeper understanding of physiological signals, a more comprehensive grasp of the current scientific landscape, and a clearer insight into emotional awareness concerns. Emotion recognition across multiple modes is particularly challenging, largely because it is hard to pinpoint features that are linked to human emotions. Maximizing the use of both audio and visual information is essential. (H) According to Zhou et al. [2], a multimodal fusion network for audio-visual emotion recognition has been proposed.

The approach begins with a feed-forward neural network that incorporates one-dimensional attention, subsequently followed by normalization. After that, a graph-based fusion block process is used to combine audio-visual information. An adaptive version of the Alternating Direction Method of Multipliers (ADMM) is used, which is based on Forward-Backward Splitting (FBP), to dynamically calculate the fusion weights of the representation vectors, thereby enhancing performance. The proposed approach, AMFBP, employs local emotion. Validation is carried out utilizing the IEMOCAP data-set solely with the audio stream. A level of accuracy of 71.40% has been established. The AFEW database and IEMOCAP are also utilized for emotion recognition in auditory-visual contexts. This method achieves the highest accuracy rates of 63.09% and 75.49%.

The Area of Emphasis in Research (AER) has received relatively less attention. Most focus is placed on identifying emotions through non-musical sounds. Understanding the impact of sound on emotional reactions may enhance the sound designer's craft. Cunningham et al [3] employ the International Affective Digital Sounds set, extracting a total of 76 features from both time and frequency domains. The characteristics are compared using the Pearson's r correlation coefficient to determine their degree of similarity. Two machine learning algorithms, namely regression and ANN, are employed in the context of emotional dimensions to utilize these features. A relatively small number of substantial correlations have been identified between the traits and the extent of traits are employed to predict emotional valence.

ANN models surpass regression models in performance, and the most effective of these networks achieve prediction accuracy rates of 64.4% for arousal and 65.4% for valence. Chen et al. [4] conducted a study involving a multimodel data set and evaluated classification results for audio, video, EMG, and EEG data. The results are derived from conventional feature extraction techniques and machine learning methodologies. Initially, a data-set comprised of 11 human subjects was compiled, incorporating six distinct emotions and a neutral category. The subsequent feature extraction process was conducted using PCA, auto-encoder, conventional network, and Mel-Frequency Cepstral Coefficients. Several emotion recognition models have been compared. The analysis demonstrates that boosting bio-sensor signals improves the accuracy of emotion classification. LSTM outperformed other models, particularly in both audio and image processing, where it achieved the most effective results.

Identifying emotions from speech patterns is a challenging yet crucial endeavor. In the context of sentiment analysis and retrieval, several techniques can be utilized to extract emotions from speech, including classification and analysis. Recent studies have employed deep learning methods to investigate the application of single-electron relaying (SER), as reported by R. A. Khalil et al. As cited in reference [5], have described deep learning techniques and conducted a review of relevant literature that utilized these methods for the purposes of speech enhancement and restoration (SER). This review scrutinizes the data-set, tackles emotional aspects, assesses the value of the contributions about SER, and highlights areas for improvement. One of the most valuable attributes of human beings lies in their capacity to understand and communicate ideas. We have undergone training and are aware of numerous emotions. The task of machine emotion recognition is complicated by the inherent subjectivity of mood, as previously observed by R. R. Choudhary et al [6] propose a system that detects and validates distinct segments of a conversation, prioritizing understanding of its semantic meaning through emotion recognition. Using deep learning techniques, including CNNs and LSTMs, they are able to classify emotional content. Furthermore, models that utilize MFCCs have been created for future use of sound data. CNN was evaluated on the RAVDESS and TESS data-sets, resulting in an accuracy of 97.1%.

Table.2 Review summary

Ref.	Type	Domain	Datasets	Methods
[1]	Review	Human-computer interface, emotional awareness problems	-	Multimodal and unimodal solutions
[2]	Implementation	Multimodal emotion recognition	IEMOCAP and AFEW dataset	FBP, FCNN, G-FBP, and AMFBP
[3]	Implementation	Audio Emotion Recognition	International Affective Digital Sounds	regression and ANN
[4]	Implementation	Posed multimodal emotional dataset and human emotion classification	multimodal dataset based on 11 human subjects	KNN, SVM, random forest, MLP, LSTM model, and CNN
[5]	Review	Emotion recognition from speech signals	-	Deep Learning
[6]	Implementation	Emotion recognition using audio	RAVDESS and TESS datasets	CNNs LSTMs, MFCCs
[7]	Implementation	Emotion Recognition in video	Wild 2017 video	Deep network transfer learning, Spatial temporal model fusion, Semi-auto reinforcement learning
[8]	Implementation	Emotion recognition in video data	fer2013 dataset	Deep learning
[9]	Implementation	Speech emotion recognition	RAVDESS, Emo-DB, and language-independent datasets	MFCC and hybrid LSTM
[10]	Implementation	Multimodal emotion recognition system	RAVDESS, and CREMA-D	3D-CNN,2D-CNN, cross-attention fusion

Ouyang et al. [7] categorize the Emotion Recognition in the Wild 2017 video data-set into seven distinct emotions: angry, sad, happy, surprised, fearful, disgusted, and neutral. This approach employs three methods to address the difficulties of emotion recognition. Feature extraction is a key application of transfer learning. Combining various networks is achieved through model fusion. Semi-automatic reinforcement learning is employed for optimization based on dynamic feedback. This method achieves an accuracy of 57.2%, surpassing the baseline of 40.47%. The accuracy of this approach is found to be 57.2%, which is better than the baseline of 40.47%.Emotion recognition using deep learning methods has the potential to yield highly promising results. Facial expressions are a vital sign for determining emotional states, according to T.S. According to Gunawan et al [8], deep learning methods are utilized to detect emotions from video recordings. Previous academic studies have relied on video data-sets to explain the process of recognition. Results obtained from mathematical simulations.

Their results are shown to them. The fer2013 data-set was employed in trials aimed at identifying depression, resulting in a 97% accuracy in the training data and a 57.4% accuracy in the testing set. The Spoken Emotion Recognizer identifies emotional cues expressed through spoken language, which are dependent on temporal factors. Traditional classifiers are outperformed by a hybrid system in the field of SER. F. According to Andayani et al. [9], a combination of LSTM networks and encoders was proposed for identifying long-term patterns in speech signals and categorizing emotions. The extracted speech features are processed using Mel Frequency Cepstral Coefficients and then fed into a hybrid Long Short-Term Memory classifier. The results indicate a significant improvement in recognition accuracy in comparison to existing models. The model achieved success rates of 75.62%, 85.55%, and 72.49% when tested on the RAVDESS, Emo-DB, and language-independent data-sets in succession. Mocanua et al. (citation [10]) have developed a multimodal emotion recognition system, integrating audio and visual elements. The proposed approach incorporates spatial, channel, and temporal attention mechanisms into both a 3D-CNN and temporal attention into a 2D-CNN to extract relevant features. The inter-modal feature is represented by a fusion of cross-attention mechanisms applied to audio and video streams.

The researchers created a classification loss that takes into account semantic relationships, incorporating a constraint that steers the attention mechanisms. This approach simulates separability within and between classification groups by utilizing relationships between various emotion categories. Studies conducted on the RAVDESS and CREMA-D data-sets yielded accuracy levels of 89.25% and 84.57%, respectively. The evaluation conducted on the RAVDESS and CREMA-D data-sets yielded accuracy figures of 89.25% and 84.57,% respectively.

3. Literature Summary

A brief overview of the methodology used for audio data analysis is described in the preceding section. This section draws attention to the key takeaways from the gathered reviews. The review summary is also presented in Table 2. According to the information provided in Table 2, of the 10 articles that were studied, two were review and survey papers. The eight remaining papers focus on the practical application of the implementation process.

The subsequent analysis centers on the recognition of emotion. These methods employ a range of data-sets, with some being accessible to the public and others gathered by the authors. Following this, several algorithms and models have been identified as being used for audio data classification and emotion recognition, with CNN and LSTM architectures being the most commonly employed. Individual models are being surpassed by hybrid models in terms of their results. Additionally, some methods utilize time and frequency domain analysis to extract features from audio signals. Employ various conventional and cutting-edge methods for categorizing the data.

4. Proposed Work

Processing audio information can be beneficial for emotional computing, which offers a cost-effective and efficient approach. Emotion recognition technology is used across a range of different applications. Applications for emotion recognition technology are numerous and varied, encompassing tasks like identifying and diagnosing mental health conditions, understanding human behavior, enabling smooth communication between humans and artificial intelligence, and creating autonomous vehicle systems. The levels of complexity in these applications differ across their various implementations. This process involves choosing suitable data prepossessing techniques, pinpointing significant features, and ensuring precise classification results. The proposed project aims to investigate several key research questions to develop an effective emotion recognition system, specifically focusing on the following:

- Different types of audio and speech features extraction techniques
- How the features are used for recognizing the emotions
- Influence of ML and deep learning techniques in emotion recognition using audio data This study aims to create an emotion recognition system, as depicted in Figure 1, that uses audio data to replicate the findings from the research question previously addressed. The proposed method relies on an audio database and uses it as a source of input. This research utilized the audio classification data-set to identify emotions, as referenced in [11]. The data-set includes separate divisions for training and validation samples. The samples are audio files encoded in

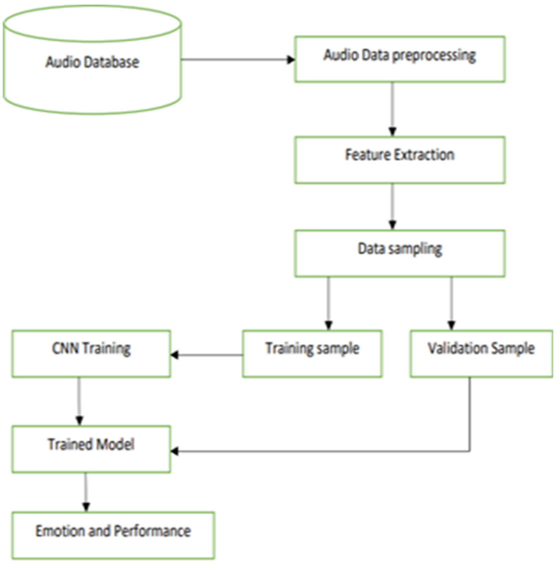

Fig.1 Proposed system for audio based emotion recognition

the MP3 format. The training set contains 5816 audio samples, in contrast to the validation set, which comprises 2492 audio files. Six emotions need to be identified. The audio samples are then prepossessed. The prepossessing involves a range of methods designed to reduce and improve the quality of data. These techniques comprise trimming, normalization, standardization, and noise removal. The primary objective of these processes is to increase the informative components of the audio signal while minimizing the non-relevant data. Audio feature extraction from both time and frequency domains must be performed to eliminate noise and equilateral the extent of time-frequency ranges. The time domain features immediately offer information about audio signals, such as energy, zero-crossing rate, and amplitude. The frequency domain feature exposes the frequency makeup of the signals, including the band energy ratio, and so on. In traditional audio data classification using machine learning, feature extraction techniques are typically implemented independently, and subsequently, machine learning algorithms are applied for classifying the extracted audio signal features. In deep learning, the models are able to extract features independently and do so more effectively than when separate features are extracted. A constitutional neural network architecture is proposed to configure for extracting audio features and then classifying them. A trained CNN model is subsequently utilized to identify emotions from audio signals. The model's performance was evaluated in terms of precision, recall, f-score, and accuracy during this process.

5. Conclusion

This paper provides an in-depth examination of existing research findings on the recognition of emotions through audio signals. Key findings from this study have been identified and scrutinized within its overall framework. These contributions concentrate primarily on emotion recognition via audio data and the application of deep learning methods. Studies have demonstrated that audio signal processing can be accomplished via two distinct approaches: by employing conventional machine learning methodologies and by utilizing deep learning methodologies.

Deep learning techniques are capable of identifying and acquiring vital characteristics directly from raw data. A proposed machine learning model aims to demonstrate the emotion recognition process, accompanied by a detailed description of the underlying methods. Following the successful development of the proposed emotion recognition model, several beneficial consequences are expected. Understanding about audio data processing and feature extraction techniques2.Understanding about the deep learning model3.Accurate emotion detection using audio signals. In the near future, the proposed model will be put into practice using Python technology, and it will be trained and tested using an audio dataset acquired from Kaggle [11]. Furthermore, this model is hosted on the Google Colab infrastructure and its outcomes will be presented momentarily.

References

[1]. S. M. S. Abdullah, S. Y. Ameen, M. A. M. sadeeq, S. R. M. Zeebaree, ―Multimodal Emotion Recognition using Deep Learning‖, Journal of Applied Science and Technology Trends Vol. 02, No. 01, pp. 73 –79 (2021)

[2]. H. Zhou, J. Du, Y. Zhang, Q. Wang, Q. F. Liu, C. H. Lee, ―Information Fusion in Attention Networks Using Adaptive and Multi-level Factorized Bilinear Pooling for Audio-visual Emotion Recognition‖, IEEE/ACM Transactions on Audio, Speech, and Language Processing, arXiv:2111.08910v1 [cs.SD] 17 Nov 2021

[3]. S. Cunningham, H. Ridley, J. Weinel, R. Picking,, Supervised machine learning for audio emotion recognition‖, Personal and Ubiquitous Computing (2021) 25:637–650

[4]. J. Chen, T. Ro, Z. Zhu, ―Emotion Recognition With Audio, Video, EEG, and EMG: A Dataset and Baseline Approaches‖, IEEE Access, VOLUME 10, 2022

[5]. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, T. Alhussain, ―Speech Emotion Recognition Using Deep Learning Techniques: A Review‖, IEEE Access VOLUME 7, 2019

[6]. R. R. Choudhary, G. Meena, K. K. Mohbey, Speech Emotion Based Sentiment Recognition using Deep Neural Networks‖, 2nd International Conference on Computational Intelligence & IoT-2021, Journal of Physics: Conference Series 2236 (2022) 012003

[7]. X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, D. Y. Huang,Audio-Visual Emotion Recognition using Deep Transfer Learning and Multiple Temporal Models‖, ICMI'17, November 13– 17, 2017, Glasgow, UK © 2017 Association for Computing Machinery

[8]. T. S. Gunawan, A. Ashraf, B. S. Riza, E. V. Haryanto, R. Rosnelly, M. Kartiwi, Z. Janin, ―Development of video-based emotion recognition using deep learning with Google Colab‖, TELKOMNIKA Telecommunication, Computing, Electronics and Control Vol. 18, No. 5, October 2020, pp. 2463~2471

[9]. F. Andayani, L. B. Theng, M. T. Tsun, C. Chua, Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files‖, IEEE Access VOLUME 10, 2022

[10]. B. Mocanua, R. Tapub, T. Zaharia b, Multimodal Emotion Recognition using Cross Modal Audio-Video Fusion with Attention and Deep Metric Learning‖, Image and Vision Computing March 21, 2023

[11]. https://www.kaggle.com/datasets/aibuzz/audio-classification-predict-the-emotions.