International Journal of Engineering Applied Sciences and Technology, 2025
Vol. 9, Issue 09, ISSN No. 2455-2143, Pages 43-47
Published Online January 2025 in IJEAST (http://www.ijeast.com)

A SURVEY ON EMOTION DETECTION USING AUDIO DATA PROCESSING AND DEEP LEARNING

Rahul Sahu, Dr. Shivangi Chandrakar, Dr. Amit Singh Rajput
Department of Microelectronics and VLSI
Chhattisgarh Swami Vivekanand Technical University
Bhilai, C.G, India.

Abstract— **The application of speech and audio data is rising day by day. In most of the applications of the speech and audio processing, the Machine Learning (ML) and Artificial Intelligence (AI) play essential role. In ML the audio data analysis is a complex task, due to a number of issues i.e. noise, variations, length, and computational cost. Therefore, audio preprocessing and features extraction is essential. These features are key highlights of the entire data in low level representation. These audio features are further utilized to train a ML algorithm and the trained ML algorithm is used for recognizing the similar audio patterns. In this paper, a review on audio data processing and deep learning based techniques are considered, which are used for identifying the emotion in audio data. Next, the review summary has been prepared and a ML based model proposed to recognize the emotions from audio data. In addition, a brief overview of the presented model has been given. Finally, the expected outcome of the model and future study plan has been discussed.**

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

I. INTRODUCTION

In this age of digitization, every second a huge quantity of data have been generated. This data comes from various sources such as from social media, education industry, entertainment agencies and more. This generated data may contain meaningful and useful insights which may be useful in various real world applications. Some of these applications may useful for society and human welfare. For example: the audio and video data analysis of the published video post in social media of a natural disaster may useful to preserve human life. However, there are more possibilities are present by analysis of the audio and video data. In recent literature, a number of audio and video data analysis methods and techniques are available, which is contributed by the different researchers and engineers. These techniques are application centric approaches and demonstrate "how machine learning and deep learning techniques can be employed on audio and video data for extracting the required insights". During investigation, it is also observed most of the models usages the multi-model feature fusion approaches, which are utilizing audio visual information to train deep learning models. But this approach is an expensive method of classification. Therefore, audio single analysis based techniques can be useful for controlling the resource use. In this presented work, the aim is to investigate the techniques available for audio data processing and investigate in terms of resource utilization. Additionally, study the role of deep learning concepts for analyzing the audio signals.

Deep learning techniques are enhanced version of Artificial Neural Networks (ANN), which includes different architectures for efficient and accurate data analysis. In recent years the deep learning prove their importance and utility in various real world applications. These techniques are capable to deal with complex, large and heterogeneous data analysis problems. These techniques can be used in medical care, image processing, EEG signal processing, predictions and others. In this presented work, the deep learning is also considered as the main area of study for analyzing the audio singles. The audio singles are time dependent information. Additionally, modelling and dependency extraction using time based data is effectively performed using Recurrent Neural Networks and its variants like LSTM and bi-LSTM. The aim of the audio signal analysis is to locate and extract the key features for identifying emotion communicated using speech data. In this section, an overview of the work covered in this paper has been discussed. The next section includes the recent development in the field of audio data analysis using the deep learning techniques. Further, the conclusion of review is described using table and its description. After that, a model is proposed and explained for future design and development. Finally, the conclusion is provided and future work has been discussed.

II. LITERATURE REVIEW

The proposed work is aimed to investigate the role of deep learning in audio data processing. Therefore, to understand the working of audio data processing some recently published articles have been collected and a reviewed in this section. In this context, by using google scholar some articles from reputed journals and conferences have been collected and most essential of them are different in this section. Based on the considered literature the collected abbreviations are also collected and given 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

Human-computer interface required to involve the end user's emotional state. This make possible to survive and use it in different application of education and medicine. Different techniques based on feelings, expressions, facial images, physiological signs, and neuroimaging is available. **S. M. S. Abdullah et al [1]** review emotion recognition signals using deep learning and compare their applications. Multimodal and unimodal solutions are studied for finding higher classification accuracy. According to finds, accuracy depends on the number of emotional classes, features, classification algorithm and dataset. This would encourage studies to understand better physiological signals of the current state of the science and its emotional awareness problems.

Multimodal emotion recognition is difficult because potential feature identification is difficult for human emotions. Therefore, full utilization of audio and visual information is essential. **H. Zhou et al [2]** propose a multimodal fusion network for audio-visual emotion recognition. First, a FCN is used with 1-D attention and normalization is performed. Next, a G-FBP is used to perform audio-visual information fusion. For improvement an adaptive AG-FBP is used to dynamically calculate the fusion weight of representation vectors. To use the local emotion, AMFBP is proposed. It is validated on the IEMOCAP dataset with only audio stream. The accuracy of 71.40% is found. Moreover, AFEW database and IEMOCAP both are used for audio-visual emotion recognition. This approach provide best accuracy of 63.09% and 75.49%.

Less attention has been observed for AER. Most of the focus is given upon emotional recognition from non-musical sound. By understanding "how sounds influence emotional response" may help to enhance the sound designer's task. **S. Cunningham et al [3]** uses the International Affective Digital Sounds set and a total of 76 features are extracted based on time and frequency domains. These features are analyzed to determine level of similarity between features using Pearson's r correlation coefficient. The features are used with two ML algorithms i.e. regression and ANN to emotional dimensions. It was found, a small number of strong correlations exist between the features and number of features to predict emotional valence. ANN perform better than regression models and the best performing networks able to provide 64.4% accurate prediction of arousal and 65.4% for valence.

J. Chen et al [4] describes a multimodal dataset and compares classification on audio, video, EMG, and EEG data. The results are given based on some baseline techniques of feature extraction and ML algorithms. First, a dataset from 11 human subjects are prepared which contains six emotions and one neutral class. Next, features extracted using PCA, auto-encoder, convolution network, and MFCC. A number of models have been compare for emotion recognition. The results show that bootstrapping the biosensor signals can increase emotion classification performance. The best results were obtained by KNN, additionally for audio and image LSTM found better.

Emotion recognition from speech signals is important but challenging. For SER many techniques can be used to extract emotions, including speech analysis and classification. Deep Learning techniques have recently been used for SER. **R. A. Khalil et al [5]** describes Deep Learning techniques and discusses some literature that utilized for SER. The review considers dataset, emotions, contributions toward SER and limitations.

The capacity to understand and communicate is one of the most valuable human abilities. We are trained and aware of different emotions. The emotion recognition is a challenging for machine due to the subjective nature of mood. **R. R. Choudhary et al [6]** proposes a system to acknowledge the sections of conversation, semantic content, using the feelings recognition. To categorize the emotional content, they employ deep learning techniques like CNNs and LSTMs. To use sound information for future use, models using MFCCs were created. It was tested on RAVDESS and TESS datasets and found CNN had a 97.1% accuracy.

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

X. Ouyang et al [7] presents Emotion Recognition in the Wild 2017 video into six emotions (angry, sad, happy, surprise, fear and disgust) and neutral. This solution utilizes three techniques to overcome the challenges of emotion recognition. Transfer learning is used for feature extraction. Model fusion is used to combine different networks. Semi-auto reinforcement learning is used for the optimization based on dynamic feedbacks. The accuracy of this approach is found 57.2%, which is better than baseline of 40.47%.

The deep learning techniques for emotion recognition can offer promising results. Facial expressions are considered as key feature to understand emotions. **T. S. Gunawan et al [8]** recognize the emotions using deep learning from the videos. The recognition process is described based on video datasets used by other scholarly works. Results obtained from models are presented with their performance. The experiment was carried out on the fer2013 dataset for depression detection, with 97% accuracy for training and 57.4% for testing set.

SER recognizes emotion signals transmitted through human speech where the emotions are depend on temporal information. However, a hybrid system performs better than traditional classifiers in SER. **F. Andayani et al [9]** proposed hybrid LSTM Network and Encoder to learn the long-term dependencies in speech signals and classify emotions. Speech features are extracted with MFCC and fed into hybrid LSTM classifier. The results indicate, it achieves a significant recognition improvement compared with existing models. The model reached 75.62%, 85.55%, and 72.49% success rate with the RAVDESS, Emo-DB, and language-independent datasets.

B. Mocanua et al [10] introduce a multimodal emotion recognition system, based on audio and visual fusion. They integrates spatial, channel and temporal attention mechanisms into a 3D-CNN and temporal attention into a 2D-CNN to capture the features. The inter-modal feature is captured with the help of an audio-video cross-attention fusion. Finally, by considering the semantic relations, they designed a classification loss based on a constraint that guides the attention mechanisms. Additionally, simulate by exploiting the relations between the emotion categories, with intra-class and inter-class separability. The evaluation performed on the RAVDESS, and CREMA-D datasets, which provide 89.25% and 84.57% accuracy.

III. LITERATURE SUMMARY

A brief overview of the method adopted for audio data analysis is provided in the previous section. This section highlights and key insights of collected review has been discussed. A summary of review is also represented using Table 2. According to the given information in table 2, among 10 of the considered articles 2 papers are based on review and survey. Additionally, 8 papers are belonging to implementation. Next, the considered subject of analysis is based on emotion recognition. In these methods different type of datasets are used among some of them are publically available and some of them are collected by authors. Next, the different algorithms and models has also been identified, which shows for audio data classification and emotion recognition the CNN and LSTM are the most popular architectures. In addition, the hybrid models are providing better results than the individual models. In addition, some of the methods utilizing the time and frequency domain analysis for extracting features from audio signals. Then perform classification using different type of classifiers new and old.

IV. PROPOSED WORK

The audio data analysis is beneficial for low cost and efficient emotional computing. The emotion recognition is used in a number of applications. The emotion recognition can be useful in a number of applications including mental disease diagnosis, human behavior understanding, human machine/robot interaction and autonomous driving systems. In implementation of these applications different complexities are exists. Such as selection of appropriate preprocessing techniques, identification of effective and descriptive features and accurate classification. Therefore, the proposed work is aimed to investigate the following research questions for effective emotion recognition system design:

1. Different type of audio and speech features extraction techniques
2. How the features are used for recognizing the emotions
3. Influence of ML and deep learning techniques in emotion recognition using audio data

With the aim of emotion recognition and simulation of above given research question an emotion recognition system using audio data is described in Figure 1. The proposed technique is relay on audio database and used as input. In this presented work the Audio classification dataset has been used for recognizing the emotions [11]. This dataset contains training and validation samples separately. The samples are audio files in “MP3” format. The training set contains 5816 audio samples and validation set contains 2492 audio files. Additionally, there are a total of six emotions to identify. Next, the audio samples are preprocessed. The preprocessing is a set of techniques aimed to minimize and enhance the quality of data. These techniques includes trimming, normalization, standardization, and noise removal. The ultimate aim of these processes is to maximize the information part of audio signal and reduces the non-informative part of data.

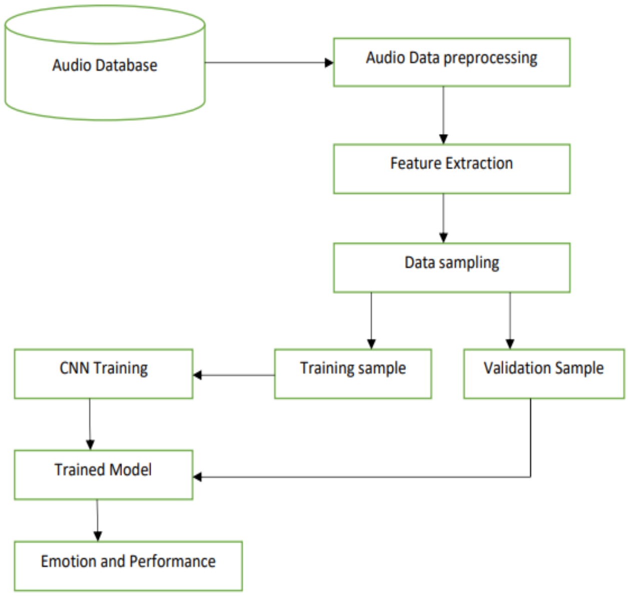

Fig. 1. Proposed system for audio based emotion recognition

The audio feature extraction from time and frequency domains is required to remove noise and balance the time-frequency ranges. The time domain feature provide immediate information about the audio signals like the energy, zero-crossing rate, and amplitude. The frequency domain feature reveals the frequency content of the signals, band energy ratio, etc. In traditional machine learning techniques for audio data classification the feature extraction techniques are separately implemented and then ML algorithms are used for classification of audio signal features. But in deep learning, the models are capable to extract the features self and can learn better then separately extracted features. Thus, a convolution neural networks (CNN) architecture is proposed to configure for extracting audio features and classify them. The trained CNN model is then used to recognize the emotions form the audio signals. During this process, the performance of the model proposed to analyze in terms of precision, recall, f-score and accuracy.

V. CONCLUSION

This paper is providing a review of existing literature for emotion recognition using audio signals. Therefore, some noteworthy contributions have been identified and discussed in this paper. These contributions are focused on emotion recognition based audio data and use of deep learning. According to the studied literature the audio signal processing can be performed in two strategies by using traditional ML approaches and by using deep learning techniques. The deep learning techniques has the ability to extract and learn the key features from the raw data. Thus, to demonstrate the process of emotion recognition using ML technique a model has been presented and an overview of the processes involved have been discussed. After successful development of the proposed emotion recognition model we expect the following fruitful results.

1. Understanding about audio data processing and feature extraction techniques
2. Understanding about the deep learning model
3. Accurate emotion detection using audio signals

In near future the proposed model is implemented using Python technology and the Audio dataset from Kaggle [11] is used for model training and validation. This model is hosted on Google Colab infrastructure additionally the results has been discussed in near future.

VI. REFERENCE

[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. Zahariab, “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.