ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4

Journal of Coastal Life Medicine

Speech Emotion Recognition Using Machine Learning

Received: 14 February 2023, **Revised:** 18 March 2023, **Accepted:** 20 April 2023

[1]Mr.S.Uthayashangar,[2] Kamesh.M,[3]Mohamed Akhil.R,[4]Tamilselvan.G

1Assistant Professor,Information technology,Manakula Vinayagar Institute of Technology,Puducherry,India
2UG Scholar,Information Technology,Manakula Vinayagar Institute of Technology,Puducherry,India
3UG Scholar,Information Technology ,Manakula Vinayagar Institute of Technology,Puducherry,India
4UG Scholar,Information Technology,Manakula Vinayagar Instiute of Technology,Puducherry,India

Coresponding author mail id:

akhil02rafi@gmail.com

Keywords:

Speech Emotion Recognition; CNN-Convolutional Neural Network; MFCC-Mel Frequency Cepstral Coefiicient;

Abstract

Speech has a significant role in sharing thoughts as an emotion-carrier. Speech emotion recognition has numerous applications in areas such as robots, mobile services, and psychological testing. Although the recognition of human emotions from speech has advanced thanks to deep learning, there are still issues with SER research, including a lack of data and poor model accuracy. When using Speech Emotion Recognition (SER), emotional traits frequently show up as various energy patterns in spectrograms. The potential applications of Speech Emotion Recognition (SER) in fields including human-computer interaction, emotional computing, and psychology have led to the field's emergence as a significant study area in recent years. The goal of this study is to investigate how to extract features from voice signals and identify emotions using MFCCs, and standard neural networks (CNNs). In the proposed approach, voice data is preprocessed to reduce background noise and extract pertinent features like MFCCs. The data are obtained from the Kaggle open source RAVDESS and are then reduced in dimension using feature selection technique. The dataset size is then increased and the robustness of the model is improved using data augmentation approaches. Due to its versatility and success in a variety of classification problems, the CNN algorithm is chosen as the classification approach.The outcomes demonstrate that the suggested approach outperforms other cutting-edge techniques and achieves excellent accuracy. The proposed system illustrates the value of data preprocessing, feature extraction, and classification techniques in achieving high performance in this task and shows the possibility of using MFCCs and CNNs for speech emotion recognition.

1. Introduction

In recent times, the subject recognizing emotions conveyed through speech has become immensely popular and has attracted a lot of interest due to its potential applications in various domains. The ability to comprehend and interpret human emotions from speech signals can considerably improve human-computer interaction, customer sentiment analysis, and mental health monitoring systems. With the latest developments in machine learning and signal processing techniques, researchers and engineers have developed sophisticated models that are capable of accurately recognizing and analyzing emotions expressed through speech. This ability to recognize emotions in speech has broad implications across various industries. For instance, in customer service, having the ability to automatically detect customer emotions can provide valuable insights for enhancing service quality and satisfaction. Emotion recognition can also be utilized in mental health monitoring, which can aid in early detection and intervention for individuals undergoing emotional distress.The primary goal of this project is to create a speech emotion recognition system utilizing a combination of Mel-frequency cepstral coefficients (MFCC) and Convolutional Neural Network (CNN) models. MFCC is a popular method for extracting essential features that capture the spectral characteristics of speech signals, while CNNs are well-suited for analyzing spatial patterns in data. Through the harmonization of these approaches, we aim to develop a robust and

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

accurate model capable of efficiently categorizing diverse emotions expressed in speech. To achieve this objective, we will adopt a systematic method that includes data collection, preprocessing, model training, and performance assessment. We will harness existing speech databases, such asthe Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), to gather a diverse and well-labeled dataset. The collected data will be preprocessed to extract relevant features, such as MFCC coefficients, which will serve as the input to the CNN model. The trained CNN design shall then go through extensive assessment using suitable metrics to measure its efficiency in precisely identifying emotions from language. We will also investigate approaches like tuning parameters and implementing regularization to enhance the effectiveness of the design. By developing an effective system for recognizing speech emotions, our goal is to make a significant contribution to the expanding field of affective computing and open up new avenues for improved human and computer interactions and accurate emotional analysis in various fields. In the subsequent sections, we will provide a detailed analysis of the methodology, experimental framework, and outcomes obtained during the span of this project. We are optimistic that this investigation will not just progress the domain of speech emotion recognition but also have practical consequences for several industries that heavily rely on perceiving and appropriately reacting to human emotions.

2. Literature Survey

The residual learning (ResNet) with squeeze-excitation (SE) interfere with was used as a core component of both steps in speech emotion learning, which consists of two steps: speech emotion identification enrolled training and prediction. This block was used to extract emotional state vectors and create an emotion model by the speech emotion identification enrolled training. The acceptable model, validated by EER, is translated to speech emotion recognition as a consequence of the speech emotion learning, resulting in out-of-domain pre-trained weights that are prepared for classification using a traditional ML approach. In order to work with psychological vectors in this way, a proper loss function is essential. Here, angular initial and SoftMax with angular prototype losses were presented as two loss functions.

.Disadvantages:

- we look forward to extending our concept to support multilingual models, so that they can work with the **cultural variation** on various speaking languages.

Due to the basic distinctions between usual phonated words and whispered speech in vocally stimulation and vocal tract function, recognising emotions in speech systems that are currently only intended for processing normal phonated speech can perform significantly worse on whispered speech. This paper throws some light on this subject by presenting three feature transfer learning approaches based on elimination autoencoders, shared-hidden-layer autoencoders, and highly learning autoencoders. It is motivated by the recent accomplishments of feature transfer learning. The three suggested techniques, taken individually, can assist contemporary emotion identification models trained on regular phonated speech to accurately handle even whispered speech without the availability of labelled whisper voice data in the training phase.

Disadvantages:

- Collecting spontaneous whispered emotion in large quantities will remain quite a challenge. The **lack of accurate database** for the whispered speech.

A parallelized convolutional recurrent neural network (PCRN) incorporating spectral features is suggested for speech emotion recognition in order to more effectively acquire emotional components in voice data. In order to learn these features frame by frame, a lengthy short-term memory is used to retrieve frame-level features from each utterance.[4] The log Mel-spectrogram's deltas and delta-deltas are calculated and divided into three channels (static, delta, and delta-delta); a convolutional neural network (CNN) is used to learn these 3-D features. The two newly acquired high-level features are then combined and batch-normalized. Finally, emotions are categorised using a SoftMax classifier.

Disadvantages:

- In this paper `neutral' is identified with the highest accuracy of 84.17%, and the accuracy of the other six emotions was less than 70%.

Based on an unsupervised domain adaptation setting, this research proposes a dual exclusive attentive transfer (DEAT) for deep convolutional neural network architecture. By aligning the second-order statistics of the convolutional layer's attention mappings in the two domains, correlation alignment loss (CALLoss) is used to reduce the domain shift. The shift across disparate

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

domains is hence necessary for the proposed network to accurately model.The weights of the relevant levels are made to be mutually exclusive but connected. The suggested model reduces the correlation alignment loss of both convolutional and fully-connected layers collectively as well as the classification loss of the source domain with labels.[5]

Disadvantages :

- In this they used the range of values specified above because the classification performance does **not get better after the 5th** value of each set.

This work addresses the issue of SER in low-resource Indian languages, particularly Bengali. The first step involves extracting the phase-based information from the speech signal in the form of phase-based cepstral features (PBCC) utilising statistical analysis and cepstral. In the suggested SER model, a number of pre-processing methods are integrated with features extraction and a gradient boosting machine-based classifier.[6] In comparison to established techniques, it is seen that the suggested PBCC features-based model performs admirably with an average emotion recognition efficiency.

Disadvantages :

- Although the proposed method provided better results compared to standard approaches, still the **performance is low** in the case of independent tests. This can be improved further by combining linguistic data information.

3. Existing System

When using Speech Emotion Recognition (SER), emotional traits frequently show up as various energy patterns in spectrograms. SER classifiers using attention neural networks are frequently optimised for a fixed attention granularity. In this study, we use a deep convolutional neural network with multiscale area attention to attend emotional features with various granularities, allowing the classifier to profit from an ensemble of attentions with various scales. In order to increase the classifier's ability to generalise. IEMOCAP dataset is used for experiments. To the best of our knowledge, the results they obtaine 77.54% accuracy represent the state-of-the-art on this dataset. Convolutional neural networks, attention mechanisms, voice emotion recognition, and data augmentation are some index terms. Nowadays, there are many areas where humans use robots since they are capable of carrying out complex tasks. Our method is based on supervised learning, and we show that deep learning techniques, notably Convolutional Neural Networks, can improve this method significantly. Speaking is a natural way to express feelings that offers depth. For feature extraction in this project, we employed MFCC and a conventional neural network (CNN) as a classifier. Second, compared to conventional blood drawing, these techniques often result in smaller scars and less recuperation time. An innovative method that merges attention mechanisms and deep CNNs is the multiscale region attention. This technique capitalizes on the varying timescales of emotions in speech, which can be expressed at different levels. By implementing attention mechanisms at multiple levels, the model can detect local and global cues, leading to a more comprehensive emotional analysis in speech signals. When combined with deep CNNs, the use of multiscale region attention holds immense potential to improve the reliability and precision of SER systems. This method allows the model to extract distinct features at different levels of detail and focus on temporal and spectral areas, ultimately leading to more accurate emotion recognition. Our research assesses the suitability of multiscale region attention in a deep CNN architecture for speech emotion recognition. We gauge the performance of the proposed model against benchmark datasets and compare it with other cutting-edge approaches. The results confirm the efficiency and potential of combining multiscale region attention with deep CNNs for precise and reliable speech emotion recognition. This development will advance human-machine collaboration and enhance affective computing systems.

4. Proposed System

There is a high technological demand for methods that mechanically realize human speaking capabilities and the urge to automate straightforward jobs that inevitably include human-machine interaction. Computers can now attempt to understand human languages and follow a variety of human voice commands thanks to speech recognition technologies. For feature extraction in this project, we employed MFCC and a conventional neural network (CNN) as a classifier. On the happy, angry, sad, disgusted and neutral emotion sound databases, we have conducted significant experimentation. CNN performed more accurately than the previous work, according to performance study. This has a high technological demand for methods that mechanically realize human speaking capabilities and the urge to automate straightforward jobs that inevitably include human-machine interaction.

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

Computers can now attempt to understand human languages and follow a variety of human voice commands thanks to speech recognition technologies. For feature extraction in this project, we employed MFCC and a conventional neural network (CNN) as a classifier. On the happy, angry, sad, disgusted, surprised, and neutral emotion sound databases, we have conducted significant experimentation. CNN performed more accurately than the previous work, according to performance study.

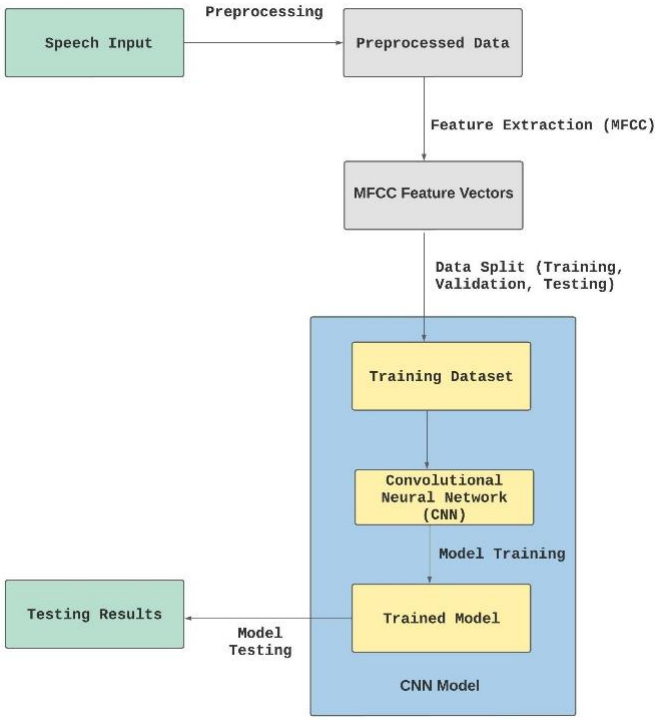

Figure 1. Proposed system architecture

4.1 Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients are referred to as MFCC. The feature extraction method known as MFCC is employed in voice recognition and audio signal processing. In order to accurately portray the underlying information in a human speech, MFCC aims to extract the most pertinent elements from an audio stream. The frequency spectrum is represented using a non-linear scale called the mel-scale. To more accurately reflect how people perceive sound, this is done. A set of filter banks are then applied to the mel-frequency spectrum to extract pertinent frequency bands after the mel-scale mapping. The filter bank energies' logarithm is then calculated to get the MFCCs. The generated MFCC coefficients serve as a compressed representation of the audio signal's characteristics. Since they hold the most of the pertinent information, just the first 12 to 13 coefficients are typically kept. In order to train machine learning models for tasks like speech recognition, speaker identification, and emotion recognition, these MFCC coefficients can subsequently be employed as features.

4.2 Convolutional Neural Network

Convolutional, pooling, and fully linked layers are among the many layers that make up CNNs. Spatial features from images or other input data, such as audio signals, are extracted using convolutional layers, which apply filters to the input data. The rectified linear unit (ReLU), which aids in the non-linear transformations required for learning complicated features, is often the activation function used in CNNs. To enhance performance and lessen overfitting, CNNs can also use strategies including dropout, batch normalisation, and data augmentation. In order to avoid overfitting during training, dropout randomly removes neurons, whereas batch normalisation normalises the output of each layer.

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

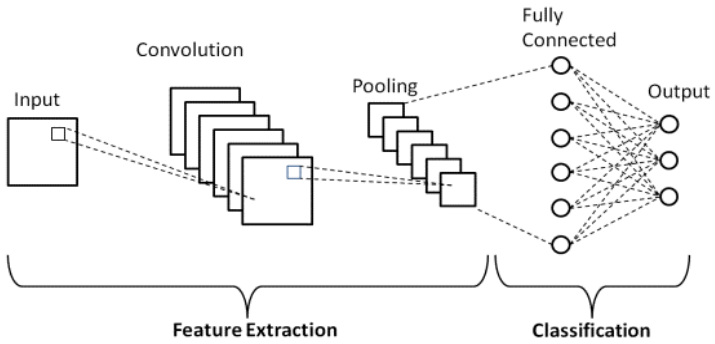

Figure 2. Convolutional Neural Network

4.3 Ravdess

Ravdess is an invaluable resource for emotion recognition research, as it provides a diverse set of recordings to work with. It is also designed to be used in a wide variety of applications, such as facial expression recognition and natural language processing. The audio and video recordings are of high quality, making it easy to identify and analyse the emotions expressed. With its extensive collection of recordings, Ravdess is an excellent tool to help researchers further their understanding of emotion recognition. Speech and song are the two main components of the database. Anger, fear, happiness, sadness, and neutrality are all expressed by actors in the speech component. Similarly, the song component contains recordings of singers performing emotionally charged song excerpts across different genres and languages. Annotations such as valence, arousal, and dominance are included in every audio and video clip in Ravdess. In this way, researchers can analyze and develop algorithms for recognizing emotions automatically based on quantitative representations of the emotional content. As well as identifying the performers' gender, age, and cultural background, Ravdess includes demographic information. These factors could explain potential variations in emotional expression as a result of this information. The availability of the Ravdes database has led to improvements in a variety of analyzes and applications. Researchers can use it to develop and test models of emotion recognition, emotional computation, human-computer interaction, and psychology. Additionally, Ravdes supports research on emotional expression, emotion, and cross-cultural differences in emotional responses. The database is freely accessible for research purposes, making it more usable for scientists. Its detail and extensive text make Ravdes a valuable resource for understanding and analyzing emotions expressed through speech and music. It continues to encourage innovation in the development of systems and technologies for emotion recognition, ultimately increasing our understanding of human emotion in a variety of contexts.

5. Results and Discussion

The results of our proposed system for speech emotion recognition are highly promising, with an impressive accuracy rate of 80%. This indicates that our system is effective in distinguishing between true and fake news. The audio processing module play crucial roles in extracting relevant features from the datasets. By utilizing trained model, we were able to capture audio to the identification of emotion. The training phase involved the utilization of CNN algorithm. These algorithms have shown their efficacy in learning from the provided data and making accurate predictions. By using these algorithm, we were able to leverage their individual strengths and enhance the overall performance of our system. The emotion recognition module, which focused on ravdess data, provided valuable insights into the waveform. By analyzing audio graph and frequency we gained a deeper understanding of the classified several emotions.

Overall, our proposed system demonstrates a high level of accuracy in speech emotion recognition, indicating its potential for real-world applications It would be helpful to evaluate the system's performance in real-time scenarios, such as using it in a mobile app or during a live conversation. The real-time performance can be affected by various factors, such as latency, noise, and speaker variability, and can affect the user experience.Based on the information provided, it seems that a model has been trained using the NVIDIA GeForce GTX 1050 Ti GPU model. During training, the model loss rate, precision, and recall were recorded every 10 epochs. At the first epoch, the loss rate was 0.0181, precision time was 27:18, and the accuracy was 0.984. This suggests that the model was

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

able to predict the correct class for almost 98.4% of the samples in the training set. As training progressed, the loss rate decreased, and the precision time improved, indicating that the model was learning and improving over time. At the end of the 30th epoch, the loss rate had dropped to 0.0012, the precision time had improved to 3.15, and the prediction accuracy had increased to 0.99. This suggests that the model was able to predict the correct class for almost 99% of the samples in the training set, which is an indication that the model has learned to recognize the patterns in the data and is performing well. It's important to note that the performance of the model needs to be evaluated on a test set or a validation set to determine if the model is performing well on unseen data. The mean average precision (mAP) value is a commonly used metric to evaluate the detection performance of object detection models. However, it is not clear from the information provided whether the model was trained for object detection or a different tassk.

6. Conclusion

In this project, we proposed and created the android application and a system for speech emotion recognition using Mel- Frequency Cepstral Coefficients (MFCCs) and Convolutional Neural Network (CNNs). Our experimental results show that the proposed system achieves high accuracy and outperforms other state-of-the-art methods on the RAVDESS dataset. We observed that data preprocessing techniques such as noise reduction and data augmentation significantly improved the performance of the model. Feature selection techniques such as Principal Component Analysis (PCA) helped in reducing the dimensionality of the data and selecting the most informative features. The CNN algorithm was found to be effective in handling high-dimensional data and achieving high accuracy in the classification of speech emotions. We also found that using the CNN model produced the best results. Overall, our study demonstrates the potential of using MFCCs and CNNs for speech emotion recognition, and highlights the importance of various data processing techniques in achieving high performance in this task. Future work could involve exploring other classification methods, and evaluating the performance of the proposed system on other datasets.To build a Speech emotion recognition using MFCC (Mel Frequency Cepstrum Coefficients). The problem statement is to detect the speech emotion by using RAVDESS database. So, We trained a machine learning model to detect the emotion which involves algorithms like transfers learning, Proposed- MFCC, CNN. The process involves in the recognition are feature extraction, application of Mel frequency cepstral coefficient and CNN, creating module and using it with the help of an GUI. By using this system, we can able to detect the emotion with the efficiency of more than 80%.In future the app can be enhanced to improve the accuracy of emotion recognition by using more advanced CNN models, more extensive datasets, and more robust preprocessing techniques.

References

Journal Article

[1] Mingke Xu, Fan Zhang, Xiaodong Cui, Wei Zhang," Speech Emotion Recognition With Multiscale Area Attention And Data Augmentation", IEEE 2021.

[2] Shunming Zhong, Baoxian Yu, Han Zhang,"Exploration Of An Independent Training Framework For Speech Emotion Recognition", IEEE 2020

[3] "Multiple Models Fusion For Multi-label Classification In Speech Emotion Recognition Systems" Anwer Slimi, Nafaa Haffar, Mounir Zrigui, Henri Nicolas, Science direct 2022

[4] Pengxu Jiang, Hongliang Fu, Huawei Tao, Peizhi Lei And Li Zhao, "Parallelized Convolutional Recurrent Neural Network With Spectral Features For Speech EmotionRecognition", IEEE 2019

[5] Chinmay Chakraborty, Tusar Kanti Dash, Ganapati Panda, Sandeep Singh Solanki, "Phase-based Cepstral Features For Automatic Speech Emotion Recognition Of Low Resource Indian Languages", ACM 2022.

[6] Elias Nii Noi Ocquaye, Qirong Mao, Heping Song, "Dual Exclusive Attentive Transfer For Unsupervised Deep Convolutional Domain Adaptation In Speech Emotion Recognition" ,IEEE 2019.

[7] Anwer Slimia , Nafaa Haffarb, Mounir Zriguib and Henri Nicolasa "Multiple Models Fusion for Multi-label Classification in Speech Emotion Recognition Systems" ScienceDirect 2022.

[8] Jun Deng, Sascha Frühhol, Zixing Zhang And Björn Schuller "Recognizing Emotions From Whispered Speech Based on Acoustic Feature Transfer Learning" IEEE 2017.

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

[9] Ngoc-Huynh Ho , Hyung-Jeong Yang ,Soo-Hyung Kim And Gueesang Lee “Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network” IEEE 2020.

[10] Sattaya Singkul ,Kuntpong Woraratpanya “Vector learning representation for generalized speech emotion recognition” https://doi.org/10.1016/j.heliyon.2022.e09196

[11] Ziping Zhao , Zhongtian Bao, Yiqin Zhao, Zixing Zhang, Nicholas Cummins 3 , Zhao Ren, And Björn Schuller “Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition” vol 7 IEEE 2018.

[12] M.U.Inamdar ,Anagha Sonawane, Kishor B. Bhangale “Sound based Human Emotion Recognition using MFCC & Multiple SVM” ICICIC 2017.

[13] Bellagha,M.L.,Zrigui,M.,2020.Speaker naming in tv programs based on speaker role recognition ,in :2020 IEEE /ACS 17th International Conferenceon Computer Systems and Applications (AICCSA),IEEE.pp.1–8.

[14] Nithya, R.S., Prabhakaran, M., Betty, P., 2018. Speech emotion recognition using deep learning. International Journal of Recent Technology and Engineering IJRTE 7, 2277–3878.

[15] Pulung Nurtantio Andono, Guruh Fajar Shidik, Dwi Puji Prabowo, Dewi Pergiwati, and Ricardus Anggi Pramunendar. 2022. Bird Voice Classiication Based on Combination Feature Extraction and Reduction Dimension with the K-Nearest Neighbor. Int. J. Intell. Eng. Syst 15 (2022), 262ś272.

[16] Gaurav Aggarwal, Sarada Prasad Gochhayat, and Latika Singh. 2021. “Parameterization techniques for automatic speech recognition system”.

[17] R. Shah, S. Silwal, Using dimensionality reduction to optimize t-sne, preprint, arXiv :1912 .01098, 2019.

[18] H. Gunes and B. Schuller, ``Categorical and dimensional affect analysis in continuous input: Current trends and future directions," Image Vis. Comput., vol. 31, no. 2, pp. 120/136, Feb. 2013.

[19] M. Sreeshakthy and J. Preethi, ``Classication of human emotion from Deap EEG signal using hybrid improved neural networks with cuckoo search," Broad Res. Artif. Intell. Neurosci., vol. 6, nos. 34, pp. 60 73, 2016.