e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

SPEECH EMOTION RECOGNITION USING CNN

Prathamesh Lal Bahadur Yadav[*1]

[*1]Department of Information Technology, B. K. Birla College (Autonomous), Kalyan, Maharashtra, India.

ABSTRACT

Convolutional Neural Networks (CNNs) have become a potent tool in the field of computer vision, transforming image classification, object detection, and other visual recognition tasks. In this research paper, we experimented with using the CNN model for speech classification. Speech Emotion Recognition (SER) is a crucial task in the field of human-computer interaction. It finds applications in areas like mental health monitoring, customer service, and human-robot interaction. SER systems use various methodologies to process and classify speech signals to detect embedded emotions. In this research, we used the Crema dataset, which comprises audio data in .wav format, to improve accuracy. Our research demonstrates the potential of CNN in Speech Emotion Recognition. It highlights how CNNs can simplify feature engineering and improve accuracy, thereby contributing to more emotionally intelligent human-computer interactions.

Keywords: CNN, Model, Speech Emotion Recognition, Accuracy.

I. INTRODUCTION

Understanding emotions is crucial as they are a fundamental part of human life, affecting our thoughts, actions, and interactions every day. While it's impossible to provide a comprehensive account of every emotion, I can help you explore some of the deeper aspects of key emotions to help you better comprehend them.

Emotions play an important role in our lives. Happiness is a cherished emotion that brings feelings of joy, contentment, and well-being. It is closely linked to the release of neurotransmitters like serotonin and dopamine in the brain. On the other hand, sadness is a universal emotion that arises in response to loss, disappointment, or difficult circumstances. It helps us process and adapt to challenging situations. Anger is a powerful emotion that can manifest in response to perceived threats or injustices, and it serves as a natural defense mechanism. Fear is a primal emotion that has evolved to protect us from danger. It triggers the "fight or flight" response and can be associated with increased heart rate, heightened alertness, and a surge of adrenaline. Love, on the other hand, is a complex emotion that can take various forms, including romantic love, familial love, and platonic love. Lastly, surprise is a momentary emotional response to unexpected events or stimuli. It often leads to heightened alertness and can be accompanied by a sense of wonder.

Deep learning is a subfield of machine learning that has gained significant attention and popularity in recent years due to its outstanding performance in various applications. It is a category of machine learning techniques inspired by the structure and function of the human brain, particularly artificial neural networks.

Convolutional Neural Networks (CNNs) are a type of deep learning model that are specifically designed to process structured grid data. They are a crucial component of modern computer vision and image processing tasks, taking inspiration from the human visual system. CNNs have become an essential technology for various applications, ranging from image and video analysis to natural language processing, and even some areas of audio processing.

II. LITERATURE REVIEW

1. Zhengwei Huang , Ming Dong , Qirong Mao, Yongzhao Zhan In their paper, the authors dive into the concept of Speech Emotion Recognition (SER), its significance, and related research. They put forward a novel SET system based on Convolutional Neural Networks (CNN) and demonstrated its superior performance compared to existing systems. The authors have achieved an accuracy rate of 8.4% higher than the previous state-of-the-art systems on a Greek speech emotion database that is publicly available.

2. Jianfeng Zhao, Xia Mao, Lijiang Chen, In this article, we will discuss speech emotion recognition using deep 1D and 2D CNN LSTM networks. We will explore two deep learning approaches for speech emotion recognition, namely, 1D CNN LSTM and 2D CNN LSTM. Both of these approaches have been proven to be superior to

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

traditional methods such as Deep Belief Network (DBN) and CNN. The 2D CNN LSTM network is the most effective, achieving an accuracy rate of 95.89% on the Berlin EmoDB database.

3. M.Shamim Hossain, Ghulam Muhammad, This article dives into the topic of audio-video emotion recognition and discusses various emotion features and fusion strategies. The authors used speech spectrograms and log mel-spectrograms for audio features and utilized different CNN models to extract several facial features. Additionally, they explored various intra-modal and cross-modal fusion techniques. The study achieved an accuracy of 65.5% on the AFEW validation set and 62.48% on the test set.

4. Michael Neumann, Ngoc Thang Vu, In this article, the use of attentive convolutional neural networks for speech emotion recognition is discussed. The article covers the authors' proposed model as well as prior research on speech emotion recognition. The authors conducted experiments using different input signal lengths, acoustic features, and types of emotion speech - improvised and scripted. The results revealed that the recognition performance is highly influenced by the type of speech data. Moreover, they achieved state-of-the-art results for improvised speech data.

5. Mustaqeem and Soonil Kwon, In this article, the focus is on speech emotion recognition (SER) using deep learning. The article highlights the challenges faced by SER, including the lack of accuracy improvement and high cost complexity of CNN architectures. To address these challenges, the authors propose a new CNN architecture that uses special strides to extract salient high-level features from spectrograms of speech signals. This method has been tested on two benchmark datasets and has achieved state-of-the-art accuracy.

III. METHODOLOGY

Dataset

As humans, we naturally express ourselves through speech. It makes sense then to use this form of communication with computer applications. Speech emotion recognition (SER) systems are methodologies that process and classify speech signals in order to detect emotions. Although SER is not a new field and has been around for over two decades, recent advancements have brought it back into the spotlight. These advancements make use of developments in all fields of computing and technology, meaning it is essential to stay up-to-date with the current methodologies and techniques that make SER possible. In this article, we have identified and discussed the different areas of SER, provided a comprehensive overview of current literature, and highlighted the challenges that still need to be addressed.

Proposed Model

Convolutional Neural Networks (CNNs) are neural network architectures that are specifically tailored to solve image-related tasks like image classification, object detection, and image segmentation. They are well-suited for these tasks because they can automatically learn and extract features from images. CNNs have the ability to analyze images in a more sophisticated way than traditional neural networks, making them a popular choice for computer vision tasks.

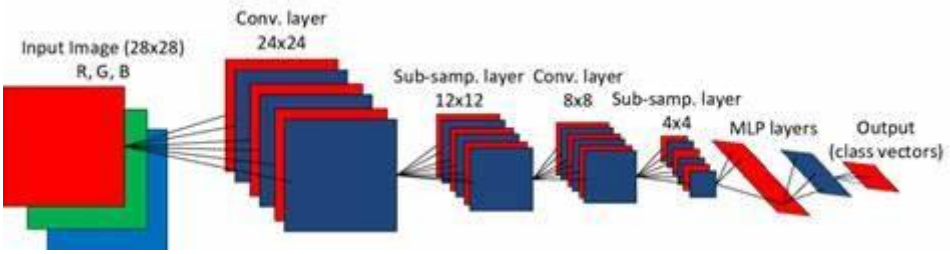

Figure 1: Architecture of Convolutional Neural Networks (CNN)

IV. EXPERIMENTAL RESULTS

Our proposed model is based on the Crema Dataset, which comprises audio recordings of six different emotions: sadness, anger, disgust, fear, happiness, and neutrality. We divided the dataset into training and testing sets in a 70-30 ratio. To validate our proposed model, we utilized a custom CNN model, which achieved an accuracy rate of 98 %.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

	precision	recall	f1-score	support
0	0.95	0.99	0.97	382
1	0.98	0.97	0.98	381
2	0.98	0.94	0.96	381
3	0.97	0.99	0.98	382
4	0.98	0.99	0.99	381
5	1.00	0.97	0.98	326
accuracy			0.98	2233
macro avg	0.98	0.98	0.98	2233
weighted avg	0.98	0.98	0.98	2233

```
val accuracy: 99.61685538291931
train accuracy: 99.41390752792358
test accuracy 97.62651141961487
```

Figure 2: CNN Model Accuracy.

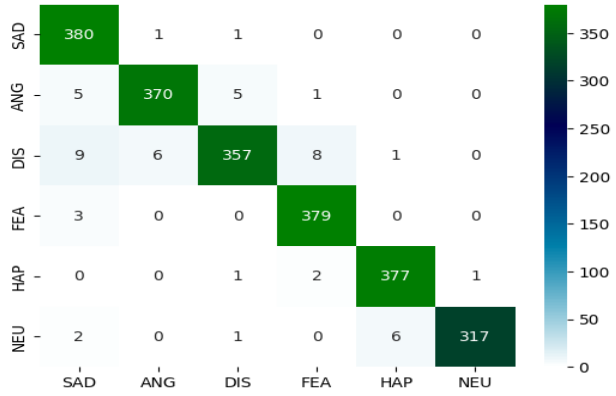

Figure 3: Error distribution of the CNN model

V. CONCLUSION

The utilization of a custom CNN for Speech Emotion Recognition has proven to be highly successful, achieving an accuracy rate of 98%. This custom Convolutional Neural Network has been specifically tailored for the complex task of recognizing emotions within speech, highlighting the potential of deep learning in analyzing audio data. With such high precision, this technology has promising applications in various fields, including healthcare and customer service, where understanding emotional cues is critical. It demonstrates the effectiveness of customization and deep learning techniques in achieving accurate results for challenging tasks.

VI. REFERENCES

[1] Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech Emotion Recognition Using CNN. Proceedings of the 22nd ACM international conference on Multimedia.

[2] Zhao, J., Mao, X., & Chen, L. (2018). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47, 312-323. https://doi.org/10.1016/j.bspc.2018.08.035

[3] M.Shamim Hossain , Ghulam Muhammad , Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data, Information Fusion(2018). https://doi.org/10.1016/j.inffus.2018.09.008

[4] Neumann, Michael & Vu, Thang. (2017). Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech.

[5] Kwon, S. (2019). A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. Sensors, 20(1), 183. https://doi.org/10.3390/s20010183