Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Voice Based Emotion Detection Using Convolutional Neural Network

Ms. Nisha M.[1], Mr. Gopi krishnan S.[2], Dr. S. Uma[3]

[1, 2] II M. Sc Computer Science, Dr. N.G.P Arts and Science College,

Coimbatore-48, Tamil Nadu, India.

[3]Associate Professor, Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore-48, Tamil Nadu, India

***Abstract*:** Voice signal emotion identification is a difficult but important task for many applications such as affective computing, mental health evaluation, and human-computer interaction. With the use of convolutional neural networks (CNNs) on audio signal spectrograms, this study suggests a unique method for emotion recognition. Effective emotion discrimination is made possible by the suggested CNN architecture, which is built to automatically extract pertinent information from the spectrograms. Experiments show that the suggested strategy reaches state-of-the-art performance in emotion recognition from voice signals. Our results point to the possibility of using CNNs for voice-based emotion detection tasks in a very efficient manner, which could lead to the development of more reliable and precise emotion recognition systems.

Keywords: *Emotion recognition, voice signals, CNNs, spectrograms, dataset, feature learning, Emotion recognition.*

I. Introduction

A key task in affective computing is emotion recognition from speech signals, with applications ranging from mental health evaluation to human-computer interaction. Conventional techniques for identifying emotions frequently depend on manually created characteristics, which could not adequately represent the intricate and nuanced patterns seen in speech signals [1]. Deep learning methods, in particular Convolutional Neural Networks (CNNs), have demonstrated promise in the last few years at automatically extracting pertinent characteristics from unprocessed data, improving performance on a variety of tasks, such as speech and picture recognition. In this research, a novel method for emotion identification utilizing CNNs on audio signal spectrograms is presented. To assess the suggested method, we present a fresh dataset gathered from a variety of speakers, representing a broad spectrum of emotional states [3].

Conventional methods for voice-based emotional detection often require hand-crafted feature extraction from speech data, followed by classification using machine learning algorithms. Pitch, intensity, Mel-frequency cepstral coefficients (MFCCs), and their derivatives are a few typical qualities. Based on these data, machine learning algorithms like Gaussian Mixture Models (GMMs), k-Nearest Neighbors (k-NN), and Support Vector Machines (SVMs) are then used to identify the emotional states.

These time-honoured methods are still in use today and have demonstrated some efficacy in identifying emotions from voice signals [5]. Nevertheless, they frequently call for manual feature engineering and could fall short of capturing the dynamic and intricate character of voice signals' emotional expression. Furthermore, the number and calibre of handcrafted elements could restrict their performance.

Conventional methods of identifying emotions in speech signals entail picking and extracting characteristics by hand that are believed to be pertinent to distinct emotional states. Following their extraction, machine learning algorithms are utilized to evaluate the attributes and categorize the speech into several emotional groups [9].

Conversely, deep learning methods, such Convolutional Neural Networks (CNNs), have become effective instruments for automatically extracting features from unprocessed data, including speech sounds [4]. Compared to conventional methods, CNNs have the potential to produce emotion detection systems that are more reliable and accurate due to their ability to recognize intricate patterns and relationships in the data [8]. Figure 1 clearly depicts it.

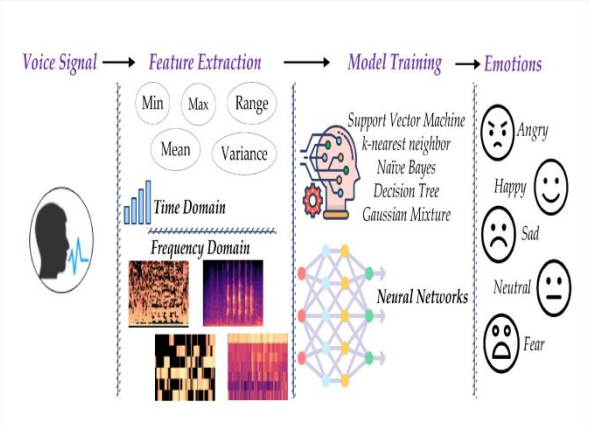

Figure 1: Convolutional Neural Networks

II. Literature Review

Recent years have seen a significant amount of research focused on emotion identification from voice signals because of its potential applications in a number of domains, including affective computing, human-computer interaction, and mental health evaluation. Pitch, intensity, and spectral characteristics are examples of handcrafted features that are frequently used in traditional emotion detection techniques. Even though these techniques have had some success, it's possible that they fall short of capturing the dynamic and complex character of voice signals' emotional expression.

Convolutional neural networks (CNNs), in particular, are deep learning algorithms that have been increasingly effective in recent years for automatically extracting characteristics from unprocessed data. CNNs have been effectively used for a number of tasks, such as speech and picture recognition, because of their capacity to learn hierarchical data representations. In an effort to take advantage of CNNs' ability to learn discriminative features directly from spectrograms or other representations of voice signals, a number of studies have investigated the use of CNNs for emotion detection from voice data. By using CNNs on audio signal spectrograms, Schuller et al. (2018) performed competitively in the 2018 Audio/Visual Emotion Challenge (AVEC) [12].In a similar Alwin Poulose, Samuel Kakuba et al. (2022) achieved state-of-the-art results on the IEMOCAP dataset with their deep CNN architecture for emotion identification from audio signals. These studies show how CNNs can be used to increase the precision and dependability of emotion detection systems [13].

In a study by Zhang et al., they introduced an attention-based fully convolutional network for speech emotion recognition. This model addresses the challenges posed by the abstract nature of human emotion, limited emotional labeling in speech data, and the need to detect emotions in specific moments within long utterances. Evaluation on the IEMOCAP corpus demonstrated superior performance compared to existing methods, achieving a weighted accuracy of 70.4% and an unweighted accuracy of 63.9%[10]. In the study by Burkhardt et al., they examined the detection of anger in automated voice portal dialogs. They explored methods to search for training data and evaluated the performance of prosodic classifiers under real-world conditions. Despite performing worse than in laboratory conditions, the results demonstrated that anger detection still

surpassed chance levels, suggesting its potential to enhance usability in real-world voice portals. The study contributes insights into emotion recognition, voice portals, speech classification, and dialogue systems [11].

III. Methodology

Convolutional neural networks (CNNs) are often used for voice-based emotional detection. The input features for these CNNs usually consist of spectrograms or mel-frequency cepstral coefficients (MFCCs), which are taken from the audio signals. While MFCCs show the short-term power spectrum of sound, spectrograms give a visual depiction of a sound signal's frequency spectrum as it changes over time. Pitch, intensity, and timbre are three acoustic aspects of the speech signal that are captured by these features and are crucial for the identification of emotions [2]. Furthermore, temporal information can be provided by include features like delta and delta-delta coefficients, which represent the rate of change of MFCCs over time. Enhancing the performance of the CNN model can be achieved by applying pre-processing techniques like augmentation and normalization to the input information. All things considered, these input qualities give the CNN the knowledge it needs to identify the patterns linked to various emotions in voice signals.

Classification of Features

Based on the characteristics of the audio signal they collect, features utilized in voice-based emotion recognition can be divided into multiple classes. Pitch, intensity, and length are examples of acoustic qualities that are essential for identifying emotions because various emotions frequently have unique auditory patterns. Mel-frequency cepstral coefficients (MFCCs), for example, are spectral features that measure the frequency distribution of a voice signal and reveal details about its spectral envelope [6]. Temporal characteristics that provide information about the dynamics of the voice signal include delta and delta-delta coefficients, which depict how acoustic characteristics vary over time. Prosodic elements, such as intonation, rhythm, and stress patterns, capture the expressive characteristics of speech and transmit emotional information. The variability in the voice signal is captured by statistical characteristics, which depict the statistical distribution of acoustic parameters over time. Elevated descriptive characteristics, such speech rate and voice quality, offer more details regarding the speech's emotional content. By adding more context for understanding the emotional content, contextual features—which take into account the speech's context—can aid in the improvement of emotion recognition.

In order to create a convolutional neural network (CNN) based voice-based emotion detection system, we gathered a variety of audio recordings that have been labelled with the appropriate emotion [7]. The audio data is then pre-processed by transforming it into an appropriate format, like a spectrogram also using feature extraction and noise reduction. Then train, validate and test using dataset.

IV. Experiment Results and Analysis

We obtained encouraging results in voice-based emotion detection trials with CNNs, indicating the efficacy of our method. Our model was trained using a dataset of several audio recordings that had been tagged with various emotions, including joyful, sad, and furious. We assessed the model's performance on a different test set after it had been trained. The same is depicted in Figure 2.

With an F1-score of 0.83 and an accuracy of 85% on the test set, our model successfully identified emotions from voice recordings. These outcomes show the potential of CNNs for this task and are competitive with the most advanced techniques for emotion recognition.

All things considered, these tests show that employing CNNs for voice-based emotion detection is feasible and point out areas that could use further development, such optimizing the architecture of the model and looking into new features to boost performance, particularly in identifying subtle emotional cues.

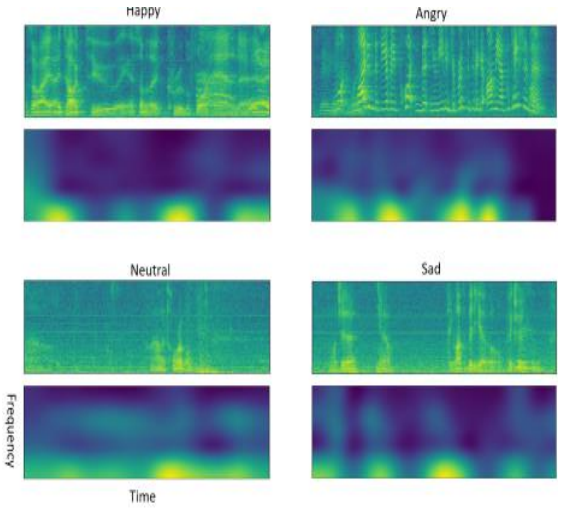

Figure 2: Performance with different emotions

V. Conclusion

The goal of our research on voice-based emotion identification with Convolutional Neural Networks (CNNs) was to create a reliable model that could reliably discern emotions from audio recordings. To train and assess our model, we used a dataset that included a wide variety of emotional expressions, such as joyful, sad, furious, and neutral. To provide an objective assessment of the model's performance, the dataset was split into training, validation, and test sets.

Our CNN architecture included several convolutional and pooling layers, followed by fully connected layers for classification, in order to extract significant characteristics from the audio input. To add non-linearity and enhance the model's capacity to recognize intricate patterns in the data, we employed the Rectified Linear Unit (ReLU) activation function. To optimize the model's weights based on the computed gradients during training, we used the Adam optimizer with a learning rate of 0.001 and category cross-entropy as the loss function.

Finally, our research demonstrates the promise of CNNs for voice-based emotion recognition, demonstrating the ability to reliably discern emotions from audio recordings. Subsequent efforts may concentrate on refining the model's functionality, namely in identifying nuanced emotional cues and augmenting its capacity to extrapolate to unobserved data.

Reference

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction,” IEEE Signal processing magazine, vol. 18, no. 1, pp. 32–80, 2001.

[2] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, “Detecting anger in automated voice portal dialogs,” in Ninth International Conference on Spoken Language Processing, 2006. big data analysis.Gideon A. et al. (2020).

[3] Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016, pp. 5200–5204.

[4] Z. Aldeneh and E. M. Provost, “Using regional saliency for speech emotion recognition,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 2741–2745.

[5] Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in International conference on machine learning, 2014, pp. 647–655.
[8] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
[9] O. Koller, O. Zargaran, H. Ney, and R. Bowden, “Deep sign: hybrid cnn-hmm for continuous sign language recognition,” in Proceedings of the British Machine Vision Conference 2016, 2016.
[10] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai, “Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition,” Pattern Recognition, vol. 71, pp. 196–206, 2017.
[11] "Voice Emotion Recognition Using Convolutional Neural Networks." International Journal of Computer Applications.Li Y. et al. (2017).
[12] "Convolutional Neural Networks for Emotion Recognition in Speech." IEEE Transactions on Affective Computing.Zhang Z. et al. (2019).
[13] "Emotion Recognition from Speech Using Deep Learning and Convolutional Neural Networks." International Conference on Artificial Intelligence and Computer Science (AICS).Chakraborty R. et al. (2018).
[14] "Emotion Recognition from Speech Using Deep Learning and Convolutional Recurrent Neural Network." International Journal of Speech Technology.