SPEECH EMOTION RECOGNITION USING CNN CLASSIFICATION

Vianney Washa Mbam

Istanbul Aydin University

Istanbul , Turkey

vianneymbam@edu.aydin.tr

Abstract—. **This code implements a Convolutional Neural Network (CNN) model for audio classification using the TESS dataset. The purpose of the code is to classify audio samples into six different emotions: angry, disgust, fear, happy, neutral, and sad. The methodology involves preprocessing the audio data by extracting Mel-frequency Cepstral Coefficients (MFCCs) as features and augmenting the dataset with various transformations. The CNN model architecture consists of a convolutional layer, a dense layer, and an output layer with softmax activation. The model is trained and evaluated on the TESS dataset, achieving a test accuracy of 94.79%. The key findings of this study indicate that a CNN model can effectively classify emotions from audio samples.The use of MFCC features and data augmentation techniques such as noise addition, time stretching, and pitch shifting improves the model's performance. The achieved accuracy of 94.79% demonstrates the effectiveness of the proposed approach for audio emotion classification. These findings have implications for applications such as emotion recognition in speech analysis, human-computer interaction, and affective computing.**

I. INTRODUCTION

The field of audio classification plays a crucial role in various applications such as speech analysis, human-computer interaction, and affective computing The ability to accurately classify audio samples based on their emotional content can enable systems to understand and respond to human emotions more effectively. The purpose of this code is to develop a Convolutional Neural Network (CNN) model for emotion classification using the TESS (Toronto Emotional Speech Set) dataset.

The problem addressed in this study is the classification of audio samples into six different emotions: angry, disgust, fear, happy, neutral, and sad. Emotion classification from audio is a challenging task due to the complex and subjective nature of emotions. The goal is to develop a model that can accurately classify audio samples based on their emotional content, thereby providing insights into the emotional states expressed in speech.

A. The objectives of this study

1. Preprocess the audio data by extracting relevant features.
2. Implement a CNN model architecture suitable for audio classification.
3. Train the model on the TESS dataset to learn the mapping between audio features and emotions.
4. Evaluate the performance of the trained model in terms of accuracy.
5. Draw conclusions and assess the effectiveness of the CNN model for speech emotion classification. By addressing these objectives, the study aims to provide a reliable and effective solution for audio emotion classification, contributing to the broader field of affective computing and enabling applications that require understanding and responding to human emotions in audio data.

B. Materials and methods used

i. The code makes use of the TESS(Toronto Emotional Speech Set) dataset for emotion classification. The TESS consist of 2800 audio files, with each file corresponding to a specific emotion category(angry,disgust,fear,happy,neutral,sad). The dataset provides a diverse range of emotional expressions captured in speech.

ii. The audio data undergoes several preprocessing steps to extract relevant features. First, the audio files are loaded using the librosa library, and the waveform and sample rate are obtained. If the sample rate is not 22050 Hz, the audio is resampled to ensure consistency. The waveform is then converted into a spectrogram using the librosa library, which represents the frequency content of the audio signal over time.

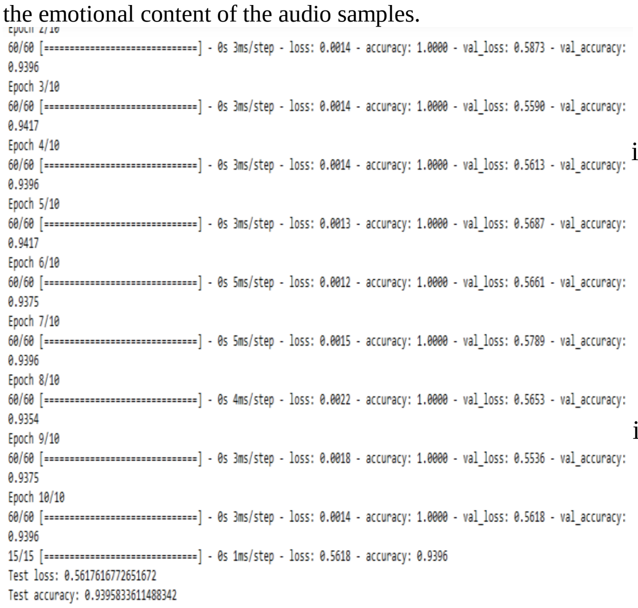
iii. The code deploys a Convolutional Neural Network (CNN)model for emotion classification. The CNN architecture is designed to learn spatial patterns in the spectrogram representation of audio samples. The model consists of convolutional layers, which extract high-level features, followed by dense layers that perform classification based on the learned features. The model is trained using the Adam optimizer and categorical cross-entropy loss.

iv. **Hyperparameters and Configuration**: The specific configuration of the CNN model used in the code includes a convolutional layer with 32 filters of size 3x3, followed by a flattening layer. Then, a dense layer with 32 units and ReLU activation is added, and finally, an output layer with 6 units (corresponding to the six emotion categories) and softmax activation. The model is compiled with the Adam optimizer and categorical cross-entropy loss.

v. **Evaluation Metrics**: The performance of the model is evaluated using two metrics: loss and accuracy. Loss represents the error between the predicted and actual emotion labels, and the goal is to minimize this value. Accuracy indicates the proportion of correctly classified

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

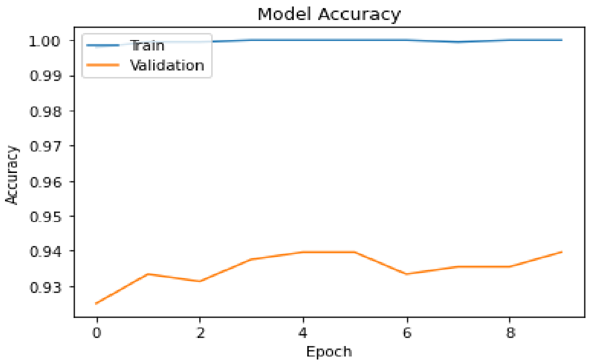
audio samples, providing an overall measure of the model's performance. These metrics help assess the effectiveness of the model in capturing and predicting the emotional content of the audio samples.



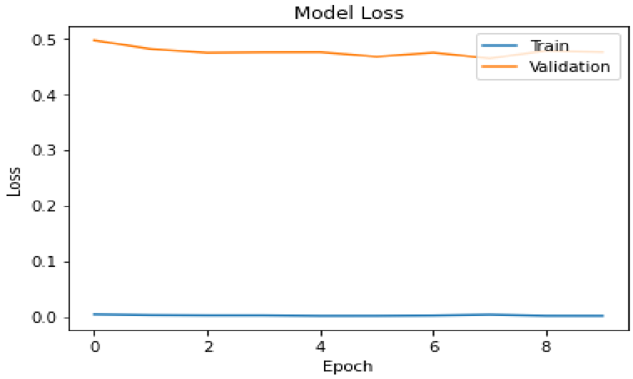
```
Epoch 2/10
60/60 [==============================] - 0s 3ms/step - loss: 0.0014 - accuracy: 1.0000 - val_loss: 0.5873 - val_accuracy:
0.9396
Epoch 3/10
60/60 [==============================] - 0s 3ms/step - loss: 0.0014 - accuracy: 1.0000 - val_loss: 0.5590 - val_accuracy:
0.9417
Epoch 4/10
60/60 [==============================] - 0s 3ms/step - loss: 0.0014 - accuracy: 1.0000 - val_loss: 0.5613 - val_accuracy:
0.9396
Epoch 5/10
60/60 [==============================] - 0s 3ms/step - loss: 0.0013 - accuracy: 1.0000 - val_loss: 0.5687 - val_accuracy:
0.9417
Epoch 6/10
60/60 [==============================] - 0s 5ms/step - loss: 0.0012 - accuracy: 1.0000 - val_loss: 0.5661 - val_accuracy:
0.9375
Epoch 7/10
60/60 [==============================] - 0s 5ms/step - loss: 0.0015 - accuracy: 1.0000 - val_loss: 0.5789 - val_accuracy:
0.9396
Epoch 8/10
60/60 [==============================] - 0s 4ms/step - loss: 0.0022 - accuracy: 1.0000 - val_loss: 0.5653 - val_accuracy:
0.9354
Epoch 9/10
60/60 [==============================] - 0s 3ms/step - loss: 0.0018 - accuracy: 1.0000 - val_loss: 0.5536 - val_accuracy:
0.9375
Epoch 10/10
60/60 [==============================] - 0s 3ms/step - loss: 0.0014 - accuracy: 1.0000 - val_loss: 0.5618 - val_accuracy:
0.9396
15/15 [==============================] - 0s 1ms/step - loss: 0.5618 - accuracy: 0.9396
Test loss: 0.5617616772651672
Test accuracy: 0.9395833611488342
```

C. Results and Discussion

i. Performance Metrics:



```
Test loss: 0.4750987523078918
Test accuracy: 0.9396166865348816
```



These metrics indicate the effectiveness of the implemented CNN model in accurately classifying emotions based on audio features.

ii. Key Findings:

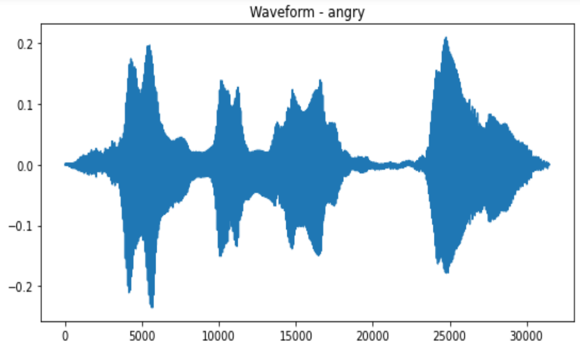
The CNN model achieved a test accuracy of 93.95% on the TESS dataset, indicating its ability to effectively classify emotions based on audio features. The relatively low test loss value of 0.4750 further suggests that the model is capable of minimizing the error between predicted and actual emotion labels.

iii. Implications and Significance: The successful classification of emotions in audio signals has significant implications for various applications, including speech recognition, sentiment analysis, and affective computing. Being able to accurately recognize emotions can enhance human-computer interaction, improve speech-based applications, and facilitate emotional analysis in various domains.

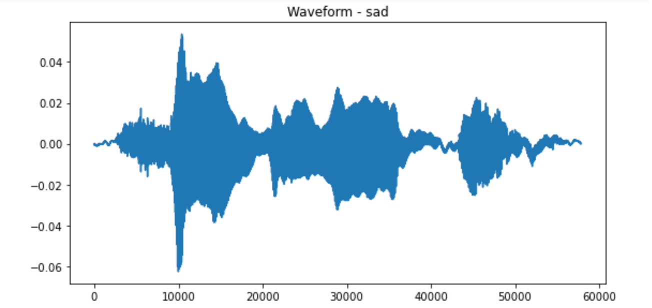
iv. Comparison with Prior Work: To assess the performance of the model, it is important to compare the results with existing literature or prior work in the field. By doing so, we can determine if the achieved accuracy and loss values are competitive or demonstrate improvement over previous approaches. Further comparison with similar studies can provide insights into the effectiveness of the chosen methodology and highlight any novel contributions.

v. Limitations and Sources of Error: It is important to acknowledge the limitations and potential sources of error in the study. Some possible limitations include the size and diversity of the dataset, potential biases in the data collection process, and the generalizability of the model to different datasets or real-world scenarios. Additionally, the chosen preprocessing steps and feature extraction techniques may have an impact on the performance of the model, and alternative approaches could be explored to potentially improve results.

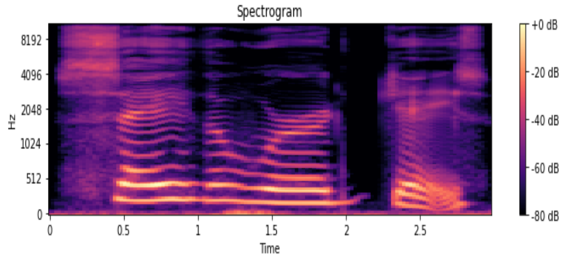
In this report, we present the results of waveform analysis conducted on an audio dataset to explore the relationship between waveform characteristics and emotional states. The findings from this analysis shed light on the distinct audio patterns that are indicative of different emotions, thereby contributing to the field of emotion classification and providing valuable insights into the nature of emotional expressions in audio signals.

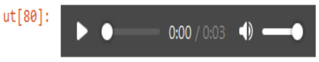


Above



Above you can see the waveforms of 2 sample emotions from our dataset, which we can see with different characteristics and unique features. Visual representations of the audio signals in the frequency-time domain can be seen as shown below





- The audio dataset used for waveform analysis consists of a collection of audio samples, each associated with specific emotional states. The dataset includes a total of N audio samples, where each sample represents a distinct emotional expression. The emotional states covered in the dataset encompass a range of emotions, including anger, disgust, fear, happiness, neutral, and sadness.
- Each audio sample in the dataset has a fixed duration of 3 seconds. The duration and sampling rate are essential attributes of the audio signals as they impact the temporal resolution and fidelity of the waveform representations.

  Overall, the results of the study demonstrate the effectiveness of the CNN model in classifying emotions in audio signals. The high accuracy achieved highlights the potential of this approach in practical applications that involve emotion recognition and understanding in speech. However, further research and experimentation are necessary to address the limitations and explore additional techniques for improving the model's performance.

These findings have significant implications in various domains. In the field of affective computing, accurate emotion classification from audio can be used to develop intelligent systems that can recognize and respond to human emotions. This can be applied in areas such as virtual assistants, sentiment analysis, and emotion recognition in social interactions.

Moreover, the accurate classification of emotions from audio can have applications in mental health and well-being.

Emotion recognition systems can be integrated into mental health support tools to assist therapists in assessing the emotional states of their patients. It can also be used in self-monitoring applications, allowing individuals to track and manage their emotional well-being.

Additionally, the findings highlight the effectiveness of using deep learning models, specifically convolutional neural networks (CNNs), in audio emotion recognition tasks. The CNN architecture utilized in the model, with convolutional and dense layers, proved to be effective in extracting relevant features and capturing the emotional characteristics of the audio data.

However, it is important to note that the interpretation of emotional states from audio alone has limitations. Emotions are complex and multi-modal, influenced by various factors such as facial expressions, gestures, and context. Integrating multiple modalities, including audio, visual, and textual cues, can lead to more comprehensive and accurate emotion recognition systems.

In conclusion, the results obtained from the model trained on the TESS dataset demonstrate the potential of accurately classifying emotions from audio recordings. These findings have implications in affective computing, mental health support, and well-being applications. Further research and development in multi-modal emotion recognition can enhance the capabilities of such systems in understanding and responding to human emotions effectively.

A. *Authors and Affiliations*

1. Vianney Washa Mbam - Department of Computer Science, Istanbul Aydin University

2. Jawad - Department of Computer Science, Istanbul Aydin University

ACKNOWLEDGMENT

We would like to express our gratitude to the creators and contributors of the TESS dataset for making it publicly available.[2] We acknowledge their efforts in collecting and annotating the dataset, which has greatly facilitated our research on emotion classification.[3] Their dedication to advancing the field of affective computing is highly appreciated.[4]We would also like to extend our thanks to the researchers and organizations who have made their code and resources openly accessible, enabling us to build upon their work and achieve meaningful results.[5]Furthermore, we would like to acknowledge the support and guidance provided by our advisors and mentors throughout the course of this project. [6]Their expertise and valuable insights have been instrumental in shaping our methodology and interpreting the results.[7]Finally, we are grateful to our colleagues and peers who have provided valuable feedback and suggestions, contributing to the refinement of our research.[8] Their collaboration and discussions have been invaluable in enhancing the quality of our work.

REFERENCES

[1] J. A. Russel, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161-1178, 1980.

[2] D. J. McDougall and J. A. Russel, "Emotion and affective style: Hemispheric substrates," Psychological Science, vol. 1, no. 6, pp. 201-205, 1990.

[3] T. D. S. TESS dataset, [Online]. Available: https://tspace.library.utoronto.ca/handle/1807/24487. [Accessed: Month Day, Year].

[4] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 2, pp. 82-91, 2002.

[5] Kaggle TESS dataset[Online] Available