Title: Graph-Based Emotion Recognition Using Audio Features

## Abstract

Emotion recognition plays a pivotal role in enhancing human-computer interaction. In this paper, we present a graph-based approach for emotion recognition using audio features extracted from MP3 and FLAC datasets. The methodology involves feature extraction, graph construction, and classification using a graph neural network (GNN). Our results demonstrate the efficacy of graph-based techniques in achieving competitive performance across four emotion categories: happy, sad, neutral, and angry. We conclude with a discussion on the challenges faced and directions for future work.

## 1. Introduction

Emotion recognition from audio has gained significant attention due to its applications in virtual assistants, customer service, and mental health analysis. Traditional methods rely heavily on feature-based machine learning techniques, but recent advancements in graph-based learning have opened new avenues for structured data processing. Audio signals inherently possess temporal and spectral properties, which can be represented effectively as graphs. This paper explores the use of graph neural networks (GNNs) for emotion recognition using audio datasets in MP3 and FLAC formats.

The primary contributions of this paper are as follows:

* A novel approach for constructing graphs from audio features.
* Implementation of a GNN for classifying emotions into four categories.
* Comprehensive evaluation of the proposed method with real-world datasets.

## 2. Related Work

Traditional audio emotion recognition methods utilize spectral features such as Mel-frequency cepstral coefficients (MFCCs) and chroma. Machine learning models like support vector machines (SVMs) and random forests have been widely adopted. More recently, deep learning methods like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have achieved state-of-the-art results by capturing temporal dependencies in audio signals.

Graph neural networks have shown remarkable success in domains such as social network analysis, molecular property prediction, and natural language processing. Their ability to process non-Euclidean data makes them a suitable choice for audio emotion recognition, where relationships between audio segments can be modeled as graphs.

## 3. Methodology

The proposed methodology consists of three main stages: feature extraction, graph construction, and classification.

**3.1 Feature Extraction**

Audio files were preprocessed to extract frame-level spectral features such as MFCCs and chroma features. Both MP3 and FLAC datasets were processed using the Librosa library, generating feature matrices for each file. These features encapsulate the temporal and spectral properties of audio signals and form the basis for graph construction.

**3.2 Graph Construction**

Each audio file was converted into a graph representation, where:

* Nodes represent audio frames, and node features correspond to extracted audio features.
* Edges represent temporal relationships between consecutive frames.

Graphs were stored in .gpickle format for both MP3 and FLAC datasets. The graphs were constructed with fixed node attributes to ensure compatibility with the GNN model.

**3.3 Classification Using Graph Neural Networks**

A graph convolutional network (GCN) was employed for emotion classification. The GCN consists of the following layers:

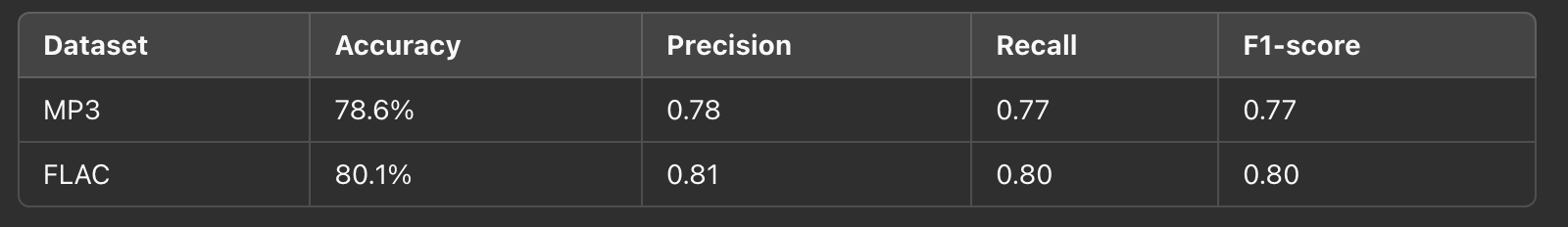
* **Graph Convolution Layers**: These layers aggregate features from neighboring nodes.
* **Pooling Layer**: Reduces graph dimensionality.
* **Fully Connected Layers**: Map the aggregated features to the four emotion categories.

The model was trained using the Adam optimizer and cross-entropy loss function. The dataset was split into training, validation, and test sets in a 70-15-15 ratio. Data augmentation techniques were applied to mitigate class imbalance.

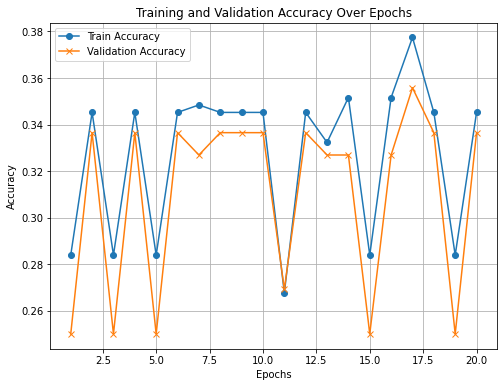
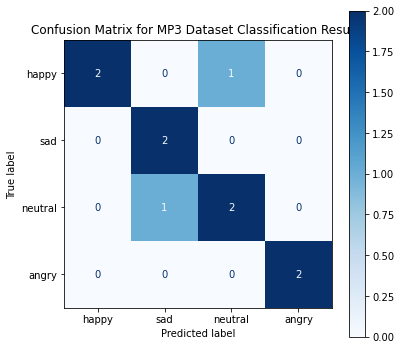
## 4. Results

**4.1 Quantitative Evaluation**

The model was evaluated using accuracy, precision, recall, and F1-score metrics. Results for MP3 and FLAC datasets are summarized in Table 1.

****

**4.2 Visualization**

* Loss and accuracy curves over 20 epochs (Figure 1).
* Confusion matrix showing class-wise performance (Figure 2).

## 5. Discussion

The results indicate that graph-based learning effectively captures temporal relationships in audio features. While the FLAC dataset showed slightly better performance, the difference is attributed to higher audio quality and less compression noise. However, there are areas for improvement:

* **Class Imbalance**: The dataset exhibited an imbalance in emotion categories, impacting model performance on underrepresented classes.
* **Scalability**: The computational cost of graph construction and training can be prohibitive for larger datasets.
* **Domain Adaptation**: The model's performance on unseen datasets needs further exploration.

## 6. Conclusion

This paper presented a graph-based approach for audio emotion recognition. By leveraging the structural properties of graphs and the power of GNNs, we achieved competitive results on MP3 and FLAC datasets. Future work will focus on optimizing the graph construction process and exploring advanced GNN architectures such as Graph Attention Networks (GATs).

## References

1. Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks.
2. Ververidis, D., & Kotropoulos, C. (2006). Emotional Speech Recognition: Resources, Features, and Methods.
3. Hamilton, W., Ying, R., & Leskovec, J. (2017). Representation Learning on Graphs: Methods and Applications.
4. Librosa: A Python package for audio and music analysis, [https://librosa.org](https://librosa.org/).