GENRE CLASSIFICATION OF HARMONIC AND PERCUSSIVE COMPONENTS OF SEPARATED SIGNALS USING DEEP EMBEDDINGS

Valentín Malpica
Universitat Pompeu Fabra
valentin.malpica01@estudiant.upf.edu

1. INTRODUCTION

In recent years, Music Information Retrieval (MIR) has become a rapidly growing field of research, with numerous applications. One of the essential tasks in MIR is music genre classification, which aims to automatically categorize music into genres based on their inherent characteristics. This study explores the performance of a music genre classification system applied to a dataset that has undergone source separation preprocessing.

The primary objective of this work is to investigate the effects of applying source separation to a dataset before performing music genre classification. We chose the GTZAN dataset for this task due to its versatility and widespread use in the MIR community. The source separation process was carried out using Demucs v.4, a state-of-the-art algorithm.

For the classification task, a two-step approach has been employed. First, we extracting features from the separated sources using some models provided by Essentia-Tensorflow. These models have proven to be effective in capturing relevant information from audio signals. Next, we used the extracted features as input to a Support Vector Machine (SVM) classifier, a popular choice for music genre classification tasks due to its robustness and ability to handle high-dimensional data.

2. DATASET

The GTZAN dataset is a widely used dataset for music genre classification tasks in the field of Music Information Retrieval (MIR). It was created by George Tzanetakis in 2002 and has since been a benchmark dataset for various MIR tasks, including genre classification. The GTZAN dataset contains 1,000 audio clips, each 30 seconds long. The audio clips are evenly distributed across 10 different musical genres, resulting in 100 audio clips per genre. The genres included in the dataset are: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, Rock. Each audio clip in the dataset is stored as a .wav file. The dataset also provides the spectrograms of the audio files, as well as two files with a large number of extracted features for each audio (one for the complete audios, the other for the 3-second version of each audio fragment). The dataset does not provide any additional metadata or annotations, such as artist or song title. Due to its simplicity and well-balanced, the GTZAN dataset serves as an excellent starting point for music genre classification tasks.

3. METHODS

In this section, the algorithms and methods used in the project to carry out the tasks of source separation and mu-

sic genre recognition are described.The first task performed is source separation, for which the Demucs v.4 algorithm has been employed. This algorithm is considered state-of-the-art due to its excellent metrics and results. Using Demucs v.4, the audios from the dataset are separated into four components or "stems": drums, bass, other instruments (others), and vocals.

For the study of the classification task, three different datasets have been chosen. Firstly, the GTZAN dataset is used without the classical music folder, as the quality of the separation of audios of this genre is not satisfactory. This is due to classical music not following the typical structures of popular and urban music found in the rest of the genres in the dataset. When applying the separation to a classical music file, three silent audios are obtained, and all the information is concentrated in the "others" stem.

In addition, two artificial datasets are generated from the combination of the stems obtained in the source separation. The first artificial dataset, called Percussive GTZAN, consists of files resulting from the sum of the drums and bass stems. The second artificial dataset, named Harmonic GTZAN, includes files that are the combination of the vocals stem and the "others" stem. This separation allows for obtaining two sets with distinct characteristics. This idea is taken from some previous work in this field. [4] [1] [2]

The feature extraction process was carried out using some of the embedding models available in the Essentia library [3]. Additional information on musical features can be found in the following publications [6] [7].These models, based on TensorFlow, provide input feature extraction and inference. They can be found on the Essentia website. VGGish is a deep VGG model trained on the AudioSet dataset, which consists of 2 million files, with the aim of extracting tags from YouTube videos. Its penultimate layer is specially designed to produce embeddings. This penultimate layer extracts 128-feature embeddings. MusiCNN [5]is another music auto-tagging model with different filter shapes aimed at the music domain. Similarly, we used the output of the penultimate layer for this model. It is trained on the well-known Million Song Dataset (MSD). EffNet is a variation of the MusiCNN model, which uses a different layer configuration compared to its original setup.

Each 30-second audio is partitioned into segments, with features being calculated for every segment. The VGG model generates a total of 32 segments and 128 features per segment. EffNet and MusiCNN, on the other hand, produce 19 and 28 segments per audio, respectively, along with a total of 200 features for each segment. We will utilize each of the three aforementioned models to process all audio files within the three datasets under examination.

A SVM classifier is applied on top of the different embeddings produced for each model-dataset combination. The Support Vector Machine (SVM) classifier is a widely-used supervised learning technique for genre classification tasks, effectively separating different music genres based on their extracted features.Two distinct classifiers have been employed: a simple SVM, previously used in Assignment 1, and an optimized SVM utilizing Bayesian search to find the best parameters for each of the three models (MusiCNN, VGGish, and EffNet).

4. RESULTS

Table 1 displays the performance comparison of three different feature extraction models (MusiCNN, VGGish, and

EffNet) using two distinct classifiers: a simple SVM and an optimized SVM. As datasets are balanced, the performance is measured in terms of accuracy, with the accuracy results shown as percentages.

From the graph, it can be observed that the optimized SVM consistently outperforms the simple SVM for all three models, achieving higher accuracy rates. Among the feature extraction models, EffNet seems to yield the best results, followed closely by MusiCNN, while VGGish appears to be the least effective model in this particular scenario. The overall trend indicates that combining the optimized SVM classifier with the MusiCNN feature extraction model leads to the highest classification accuracy.

	Normal	Percus.	Harmon.
Musicnn (SVM1)	0.90%	0.78%	0.80%
Musicnn (SVMopt)	0.94%	0.83%	0.83%
VGG (SVM1)	0.78%	0.60%	0.70%
VGG (SVMopt)	0.77%	0.57%	0.68%
EffNet (SVM1)	0.95%	0.88%	0.86%
EffNet (SVMopt)	0.98%	0.88%	0.86%

Table 1. Classification accuracy for SVM1 and SVM-opt models of the 3 datasets

The best results from the SVM classifiers have been selected and are displayed in the following figure Figure 1. In the figure, we can observe the distribution of classification accuracy for each different dataset, namely Normal, Percussive, and Harmonic. It is evident that the classifiers exhibit varying performance depending on the dataset, with Normal dataset achieving higher accuracies than the others. The figure provides valuable insights into the effectiveness of the chosen datsets, feature extraction methods and classifiers when applied to different aspects of the audio signals.

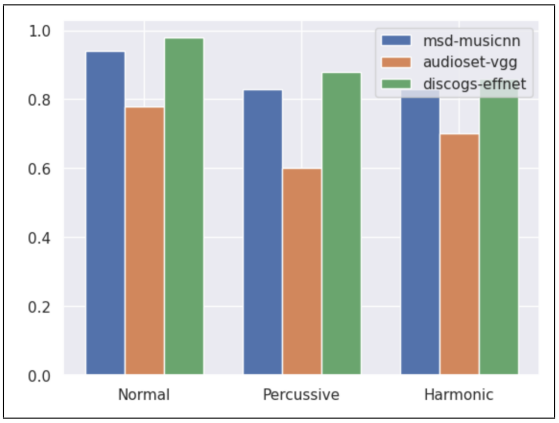

Figure 1. Classification Accuracy Distribution.

5. CONCLUSSIONS

The combination of feature extraction using Essentia models and SVM classifiers has generally produced very good results, indicating that this approach is a strong strategy for music genre classification. In particular, the musically-motivated models (both Vanilla MusiCNN and its variation, EffNet) outperform the VGG-based model, especially when working with the percussive and harmonic components separately. This highlights the effectiveness of using domain-specific neural networks for improved music genre classification across various data representations.

There are differences in accuracy when comparing the results of the Percussive or Harmonic datasets to the complete dataset. However, these differences are not very pronounced, which leads us to conclude that it is indeed possible to classify the genre of an audio file by knowing some of its components and not necessarily the entirety of the audio.

This finding demonstrates the potential for further research and development of more specialized music genre recognition systems that focus on specific aspects of audio data.

6. REFERENCES

[1] B. Kostek A. Rosner. Automatic music genre classification based on musical instrument track separation. 01 2017.

[2] B. Kostek A. Rosner, B. Schuller. Classification of music genres based on music separation into harmonic and drum components. 2014.

[3] Serra X. Alonso-Jiménez P, Bogdanov D. Deep embeddings with essentia models. 2020.

[4] Paraskevi S. Lampropoulou Aristomenis S. Lampropoulos and George A. Tsihrintzis. Music genre classification based on ensemble of signals produced by source separation methods. 2010.

[5] Serra X Pons J. musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*, 2019.

[6] Anupam B Yeshwant S. Robustness of musical features on deep learning models for music genre classification. *Expert Systems with Applications*, pages 1–9, 08 2022.

[7] Zhiqiang Zheng, Yuexian Chen, Jiaqi Li, Yuhang Li, Rui Li, and Chengzhong Li. Music genre classification based on features extracted from convolutional neural networks. *Applied Sciences*, 13(3):1476, 2023.