An interplay between genre and emotion prediction in music: a study in the Emotify dataset

Leonardo Vilela de Abreu Silva Pereira [1*], **Tiago Fernandes Tavares**[1 2]

[1]School of Electrical and Computer Engineering – University of Campinas
Av. Albert Einstein, 400, Campinas, SP

[2]Interdisciplinary Nucleus for Sound Studies
Rua da Reitoria, 82, Campinas, SP

l220198@dac.unicamp.br

Abstract. *Automatic classification problems are common in the music information retrieval domain. Among those we can find the automatic identification of music genre and music mood as frequently approached problems. The labels related to genre and mood are both generated by humans, according to subjective experiences related to each individual's growth and development, that is, each person attributes different meanings to genre and mood labels. However, because both genre and mood arise from a similar process related to the social surroundings of an individual, we hypothesize that they are somehow related. In this study, we present experiments performed in the Emotify dataset, which comprises audio data and genre and mood-related tags for several pieces. We show that we can predict genre from audio data with a high accuracy; however, we consistently obtained low accuracy to predict mood tags. Additionally, we tried to use mood tags to predict genre, and also obtained a low accuracy. An analysis of the feature space reveals that our features are more related to genre than to mood, which explains the results from a linear algebra viewpoint. However, we still cannot find a music-related explanation to this difference.*

1. Introduction

Each person relates differently to specific music pieces, that is, the same piece can trigger cathartic euphoria for some, while provoking overwhelming negative sensations for others. Sensations related to musical experiences are often regarded to as emotions and feelings, but music is also related to a sense of belonging to particular social groups, whose identity merges to that of its musical aesthetics. This dialogue foster a process in which musical choices in instrumentation, rhythms, and timbres often give rise to music genres [1].

The idea that some emotions are inherent to music were common in Europe around the XVII Century, and gave origin to the Doctrine of Affects, which directly relates particular musical structures to corresponding affects. This idea was later contested by Hanslick [2] in the XIX Century, who proposed decoupling the ideas of beautiful and the feelings related to music. Hanslick's [2] argument is that music can trigger emotions in individuals, but each one's experience is subjective and particular.

This means that music does not carry feelings by itself, the same way that spoken words do not carry feelings by themselves. Rather, feelings emerge within the individual's brains, and they are highly dependent on each one's growth and life story.

Nevertheless, humans learn adequate forms for expressing their emotions during their childhood [3]. This process is intertwined with the development of each person's communication skills, which allows expressing emotions (inner states) as feelings (externalized states). Thus, feelings are constructions that vary for each individual, but also tend to be somewhat uniform within a specific culture or ethnicity.

Like feelings and their expression, music and the process of making and listening to music are also social constructs. The responses, perceptions and uses of music are learned by individuals based on the social group they belong to [1]. With time, these identities can spread beyond small communities and become well-known, commercial music genres [4].

Feelings and music genres come from similar processes. They are learned by individuals in a dialogue with their communities, and some of their aspects can become more common, possibly being considered a part of the communities themselves.

Both feelings and musical genres can be important pieces of information about music, allowing cataloguing, analyzing, and searching within databases. For this reason, the automatic identification of feelings and of musical genres have both been broadly studied in Music Information Retrieval (MIR).

Automatic music genre identification is an MIR task that has been studied for decades now [5, 6]. It works by calculating features from audio signals, thus mapping audio files into a $\mathbb{R}^N$ vector space. Then, classifiers can use labeled data to find an optimal class-related partition within this vector space.

Music-related emotions can be predicted using a similar process, in which audio is mapped into a $\mathbb{R}^N$ vector space [7]. Interestingly, this mapping can be performed using the same features used for music genre classification [5], and classifiers use emotion-related labels to partition the vector space. This further indicates a similarity between genre- and affective-related labels in music.

*Supported by CNPq.

The correlation between genre and affective labels has been observed in work by Aljanaki et al. [8]. They observed that some music genres are more often related to specific affective labels. However, this does not happen for all affective labels.

In this study, we investigate this correlation from a machine learning point of view. More specifically, we investigate whether affect-related predictions can be useful to predict genre, and vice-versa. The goal of this study is to understand whether, and to what extent, music genres and music affects are simply different perspectives of the same phenomenon.

An important aspect of music labels is that humans often disagree among themselves regarding the genre [5], emotion [8], and even the subject [9] of songs. For this reason, we do not aim at achieving high prediction accuracies. Instead, we focus on finding explanations about the behavior of classifiers and feature spaces.

In the next section, we show the data and methods used in this investigation.

2. Methods

Currently, there are many readily-available datasets for music genre identification and music emotion recognition, as well as a myriad of solutions for these tasks. In this work, we did not focus on developing a new dataset or a new method; rather, we aimed at gathering evidences on the interplay between the genre- and affect-related labels from a machine learning perspective. We discuss the dataset in Section 2.1 and the experiments in Section 2.2.

2.1. Dataset

In this work, we used the Emotify dataser [8]. It contains 400 pieces gathered from the Magnatune dataset. Music genre labels were assigned by the recording companies, generating a balanced dataset with Rock, Pop, Classical, and Electronic genres.

Each piece was rated in a user study in which users assigned affective labels to songs. These labels came from the Geneva Emotional Music Scale (GEMS) model [10], which consists of assigning a binary value to 9 categories: Amazement, Solemnity, Tenderness, Nostalgia, Calmness, Power, Joyful activation, Tension and Sadness.

We further processed the dataset in order to find single tags for each song, as follows. First, we calculated $\mu_{s,k}$, which is the number of users that rated song s as positive for emotion k divided by the total number of ratings for song s. Then, we calculated the median m_k of $\mu_{s,k}$ along all songs for emotion k. Finally, we assigned the value 1 for emotion k in song s if $\mu_{s,k} > m_k$, and 0 otherwise.

2.2. Experiments

We performed a series of experiments using the Emotify dataset. First, we used a classification process to predict genre and affective labels directly from audio signals, as shown in Section 2.3. Then, we experimented using the results of the affective label predictions as feature vectors in the problem of identifying music genre, as discussed in Section 2.5.

2.3. Genre and emotion prediction from audio

The audio-based classification used a pre-trained VGG19 neural network [11]. The features generated by the VGG19 were yielded to a simple pipeline comprised of a scaler and a K-Nearest-Neighbors classifier. In our experiments, we used $k = 5$ neighbors. We randomly used 70% of our data for model fitting and 30% for testing.

The PCA projection of our testing data, shown in Figure 1, highlights that the VGG19 was able to find features that strongly separate music genre within this dataset. Unsurprisingly, the prediction accuracy for genre in the testing set was 100%, as displayed in the confusion matrix shown in Figure 2.

2.4. Experiment I - Audio input predicts genre

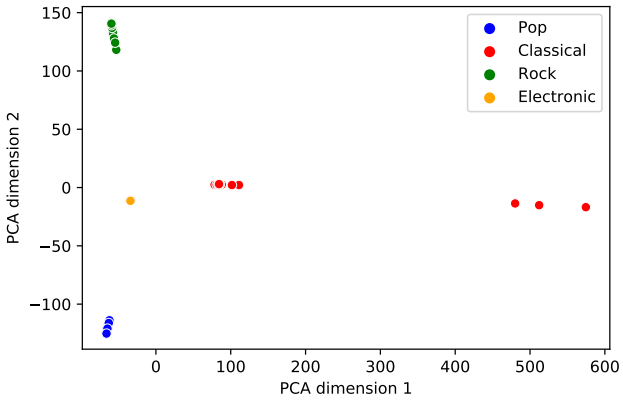

Figure 1: Audio feature space PCA projection colored by ground-truth genre labels.

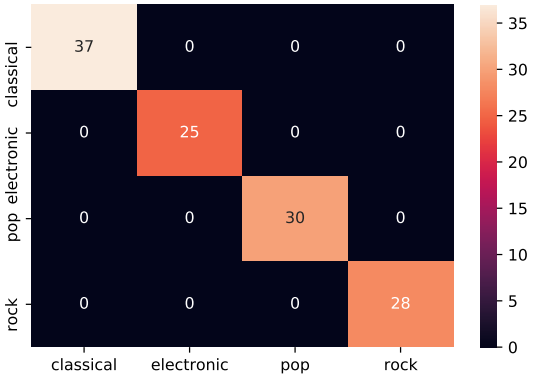

Figure 2: Confusion matrix for audio input, genre output.

We highlight that this does not mean that VGG19 is an ideal audio feature extractor. Rather, these results can be a reflection of characteristics that are specific of the Emotify dataset.

Using the same experimental setup, we conducted experiments for each one of the affective labels. The accuracy for each label is shown in Table 1. As we can see, the prediction accuracy is significantly lower than in the case of genre identification.

Table 1: Accuracy for emotion prediction using audio input.

Emotion	Accuracy
Amazement	0.52
Solennity	0.67
Tenderness	0.67
Nostalgia	0.63
Calmness	0.57
Power	0.62
Joyful Activation	0.52
Tension	0.63
Sadness	0.43

Figure 3 shows the confusion matrices related to each one of these emotions. They further confirm that predicting affective labels in this dataset is a harder task for machine learning, when compared to genre prediction.

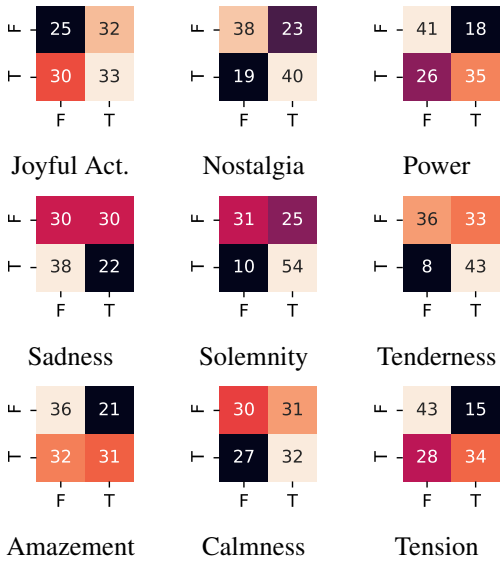

Figure 3: Confusion matrices for each emotion in the emotion prediction from audio experiment.

Although genres and affective labels come from a dialogue between individuals and their societal surroundings [3, 1], and that similar feature sets have been used to predict both genre and affective labels in the past [5, 7], we observed that the VGG19 features are much more effective to predict genre than to predict emotion. We add that previous analysis in the Emotify dataset [8] have encountered some correlations between the most frequently chosen emotion tags and music genres. However, our system had an obviously different performance when dealing with genre and when dealing with affective labels.

In the next experiments, we use GEMS ratings as feature vectors to predict genre.

2.5. Genre prediction from emotion space

The experiments shown in this section use GEMS ratings to predict music genre. We performed two of such experiments: the first used the ratings estimated in the data preprocessing stage discussed in Section 2.1, and the second used the estimated probabilities yielded by the KNN classifier discussed in Section 2.3. These values were yielded to a further KNN classifier that predicts genre-related labels, and, as with the previous experiment, we used $k = 5$ neighbors and a 70%/30% train/test data split.

The dataset GEMS ratings spans an interesting vector space, whose PCA projection is shown in Figure 4. As we can see, although it seems that each genre has a greater tendency of occupying a different region in the vector space, there is a great overlap between the regions. It is hard to determine if this is a characteristic of music genres or if it is a specific bias of this dataset.

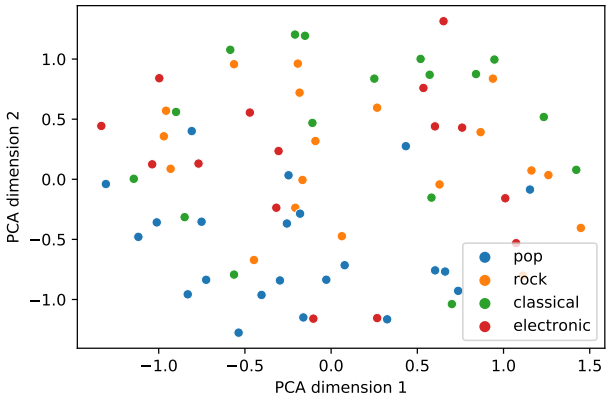

Figure 4: PCA projection of vector space spanned by GEMS ratings in the dataset.

The prediction experiment results, shown in Figure 5, indicate a greater accuracy for the Pop genre. This can be linked to the predominance of Pop songs in the lower left of the PCA projection. Likewise, this can be either a characteristic of music genres or of this specific dataset.

When we span the emotion vector space using prediction probabilities from the audio-based classifier, we have a somewhat different result. The space spanned by these predictions, shown in Figure 6, indicates an even higher overlap between the regions occupied by each genre when compared to the previous result. The confusion matrix for this experiment, shown in Figure 7, indicates a high tendency of polarization towards the Classical genre.

3. Discussion

In this work, we experiment on the prediction of genre and emotion labels from audio. After that, we try to use emotion-related labels to span a feature space that allows to predict genre. We perform experiments in the Emotify

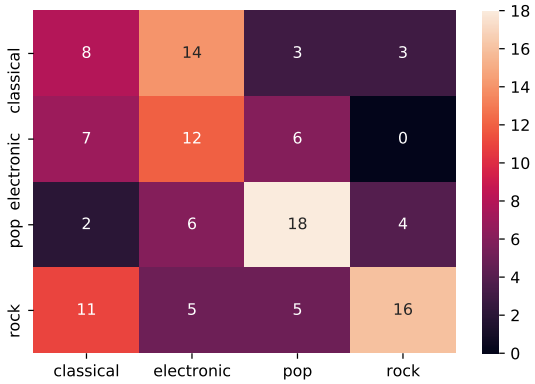

Figure 5: Confusion matrix for genre prediction using GEMS ratings.

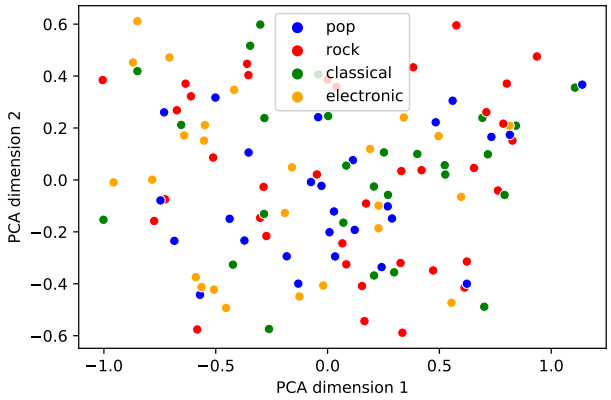

Figure 6: PCA projection of vector space spanned by GEMS ratings predicted from audio.

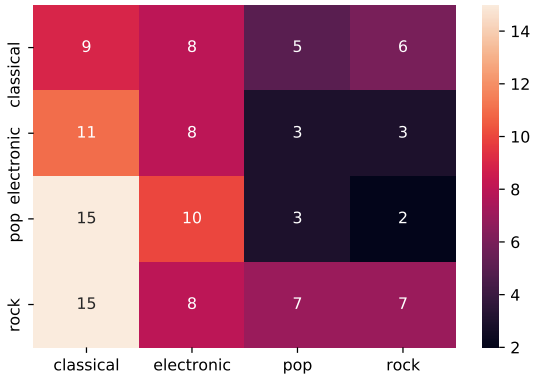

Figure 7: Confusion matrix for genre prediction using GEMS ratings predicted from audio.

dataset, which contains both genre and emotion-related labels.

We used the idea that both music genres and music-related emotions come from a dialogue between the individuals and their surrounding society. Music genres are related to the inter-social affective relationships that arise among individuals, and emotions are, as well, related to this shared individual and social identity. This idea is corroborated by studies in psychology [1] and neuroscience [3].

However, we observe that the prediction accuracy for music genre is evidently higher than that for emotion prediction. This means that, at least for our dataset, genre and emotion labels are not different viewpoints of the same phenomenon. Rather, they seem to derive from different underlying processes.

One important characteristic of the dataset is that genre labels were provided by recording companies, that is, they have a meaning related to their commercialization. On the other hand, the emotion-related tags were provided by humans, thus are related to subjective experiences [8]. Also, they use a specific model – the GEMS scale – which can also cause distortions that impact the prediction accuracy.

This accuracy, albeit relatively low when compared to genre prediction, do not mean that the experiment failed. On the contrary, they can be a result of the subjectivity of the emotional experience when listening to music. Such subjective causes humans to disagree among themselves on the GEMS ratings [8], henceforth machine learning algorithms cannot possibly achieve a high accuracy, as this would mean a super-human capacity of predicting emotions, that is, an obvious algorithmic bias.

Next, we used emotion ratings to predict music genre. This experiment had a visibly lower accuracy when compared to genre prediction from audio features. This can be due to a series of elements.

One possibility is that genre and emotion are inherently uncorrelated. Although this was observed in our dataset to some extent, it seems clear that some music genres tend to have associated affective meanings – for example, Punk music is commonly associated with teenage anger, whereas Reggae brings a much calmer mood. However, it can be the case that the procedure to produce the genre and emotion labels are uncorrelated to the social and cultural meanings often attributed to them.

Another important aspect is that the audio feature space was spanned using a VGG19, which produces vectors with around 10^7 features, while the emotion space uses 9 dimensions. Hence, the VGG19 is projecting audio excerpts into a very high-dimensional vector space, in which we observed a high separability for musical genres. However, the VGG19 was trained to perform image classifications, that is, we are using a transfer-learning technique.

This could indicate that the VGG19 features are unrelated to genre-specific musical style, and are somehow exploiting audio characteristics that derive from typical mixing and mastering processes related to each genre. This means that the high accuracy achieved in the genre classification problem could be due to properties of high-dimensional spaces and some coincidences within the

dataset. These conditions disappear when we use a low (9) dimensional space to represent tracks, hence leading to a lower accuracy.

Such reasoning brings forward the importance of the PCA projections. In Figure 4, for example, we can see that music genres overlap in the vector space spanned by GEMS ratings. This projection indicates that the prediction accuracy was not harmed necessarily by the low amount of data; rather, the genre sets are inherently correlated and it can be hard to separate between them.

Nevertheless, we observe in Figure 4 that the overlap between genres is not along the whole emotional space. Also, each genre seems to have a slightly different median in the emotional space. Consequently, we speculate that although music genres are somehow related to affective values, there is enough space within each music genre to compose music with significantly different affective value.

Next, we present concluding remarks.

4. Conclusions

In this article, we compared the prediction of genre and of emotions in music from audio using a transfer-learning technique. We also investigate the prediction of genre from a feature space spanned by emotion-related ratings.

We obtained a high accuracy when predicting music genre from audio features. We believe that this result is due to properties of high dimensional spaces and production biases (mixing and mastering techniques) in the dataset.

Also, we identified that musical genres are located in overlapping regions within the emotional-spanned feature space. This means that, regardless of the dataset size, there can be music from different genres that are related to the same emotion.

Lastly, we highlight that this work does not aim to achieve a higher prediction accuracy. Rather, we propose an analysis aimed at explaining why the prediction methods perform the way they do. We provide this explanation under the light that both genre and emotion labels are subjective.

There are many possible steps to improve our ability to explain the behavior of classifiers in the audio domain. We anticipate that using more concrete concepts, such as key or tempo, as support could lead to results that correlate more to musicological concepts. Such idea poses a vast field for future work.

References

[1] B. Ilari, "Música, comportamento social e relações interpessoais," *Psicologia em Estudo*, vol. 11, no. 1, pp. 191–198, Apr. 2006.

[2] E. Hanslick and G. Payzant, *On the musically beautiful : a contribution towards the revision of the aesthetics of music / Eduard Hanslick ; translated and edited by Geoffrey Payzant*. Hackett Pub. Co Indianapolis, Ind, 1986.

[3] A. Damasio and G. B. Carvalho, "The nature of feelings: evolutionary and neurobiological origins," *Nature Reviews Neuroscience*, vol. 14, no. 2, pp. 143–152, Feb. 2013. [Online]. Available: http://www.nature.com/articles/nrn3403

[4] J. C. Lena, *Banding Together: How Communities Create Genres in Popular Music*. Princeton University Press, 2012.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[6] J. H. Foleiss and T. F. Tavares, "Texture selection for automatic music genre classification," *Applied Soft Computing*, p. 106127, feb 2020.

[7] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2004, pp. V–705.

[8] A. Aljanaki, F. Wiering, and R. Veltkamp, "Collecting annotations for induced musical emotion via online game with a purpose emotify," UU BETA ICS Departement Informatica, Tech. Rep., 01 2014.

[9] A. Dalmora and T. F. Tavares, "Identifying narrative contexts in brazilian popular music lyrics using sparse topic models: A comparison between human-based and machine-based classification," in *XVII Brazilian Symposium on Computer Music*, São João del Rei, MG, Brazil, sep 2019.

[10] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement." *Emotion*, vol. 8, no. 4, pp. 494–521, 2008. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/1528-3542.8.4.494

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.