Emotion from music spectrograms

Matteo Cerutti
Politecnico di Torino
s265476@studenti.polito.it

Antonio Santoro
Politecnico di Torino
s264014@studenti.polito.it

Marco Testa
Politecnico di Torino
s265861@studenti.polito.it

Abstract

1. Introduction

Nowadays people need to have the possibility to select music and make playlists based on their mood. Many music platforms feature different music playlists made by hand that include popular and commercial songs aiming to maximise ratings. One of the most used feature on these platforms is to create playlists similar to other ones, the point is that all the songs that will be included are selected on the ”similarity”. Since the intention was to stay inside the computer vision domain, we have to treat audio files as images so the first idea was to exploit spectrograms. After some researches, we found that our idea was applied to classify song genres, therefore starting from the article of Piotr Kozakowski and Bartosz Michalak [2], we adapted their work to our objective. The interest is to train a neural network on different audio speeches that represent different human emotions, extract features and try to see whatever those peculiarities can be matched from music. Amiriparian *et al*. [1] showed that processing spectrograms into networks characterized by a different depth the result will change. This report presents results obtained from three networks, ResNet152, VGG11 and GoogLeNet, trained on the RAVDESS Emotional song audio dataset [3] and tested on the CAL500 dataset [4].

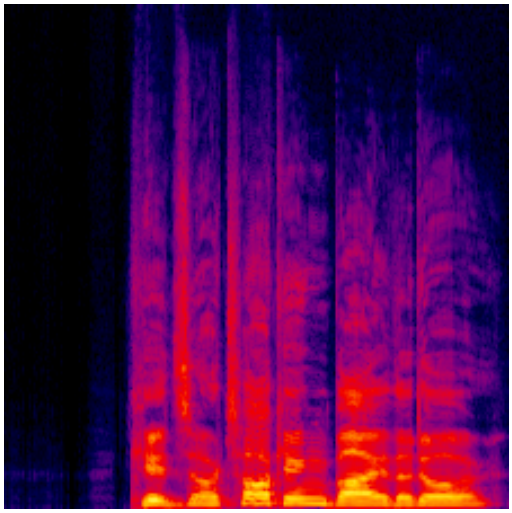

Figure 1. Training spectrogram sample.

2. Data preparation

2.1. Training dataset

The RAVDESS Emotional song audio consists of 1012 files of actors singing four seconds in a neutral North American accent. The portion used for this work includes calm, happy, sad, angry, and fearful emotions, each vocal is produced at two levels of emotional intensity, normal and strong.

Files are provided as .wav (16bit, 48kHz) that need to be converted into a raw spectrogram. For the purpose ”SoX (Sound eXchange) sound processing utilities” has been used. This tool can process audio files and do things like trimming or filters frequencies. Spectrograms for the training dataset have been generated to fit the input size of the three networks, furthermore, to cope with the limited size of the dataset, augmentation has been applied like random grey colorizing, brightness, contrast and hue variations. Amiriparian *et al*. [1] showed that using different shade of colour could exhibit different outcomes but, since the hue transformation made by the PyTorch framework has a not negligible impact on the brightness of the image and the goal is to not alter any information on the spectrogram, the best choice was to stick with the original shades, hence playing with the contrast and with a monochrome image could have a positive effect on capturing some features.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

2.2. Test dataset

3. Training phase

4. Testing

5. Conclusions

5.1. Miscellaneous

Compare the following:
`$conf_a$` $conf_a$
`$\mathit{conf}_a$` conf_a
See The TEXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

5.2. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

5.3. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LATEX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

References

[1] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Bjorn Schuller. Snore sound classification using image-based deep spectrum features. 2017.

[2] Piotr Kozakowski and Bartosz Michalak. Music genre recognition. 2016. `http://deepsound.io/music_genre_recognition.html`.

[3] Livingstone SR and Russo FA. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. plos one 13(5): e0196391, 2018. `https://doi.org/10.1371/journal.pone.0196391`.

[4] Turnbull, Douglas, Barrington, Luke, Torres, David, Lanckriet, and Gert. Semantic annotation and retrieval of music and sound effects. *Audio, Speech and Language Processing, IEEE Transactions on*, 16(2):467–476, 2008.

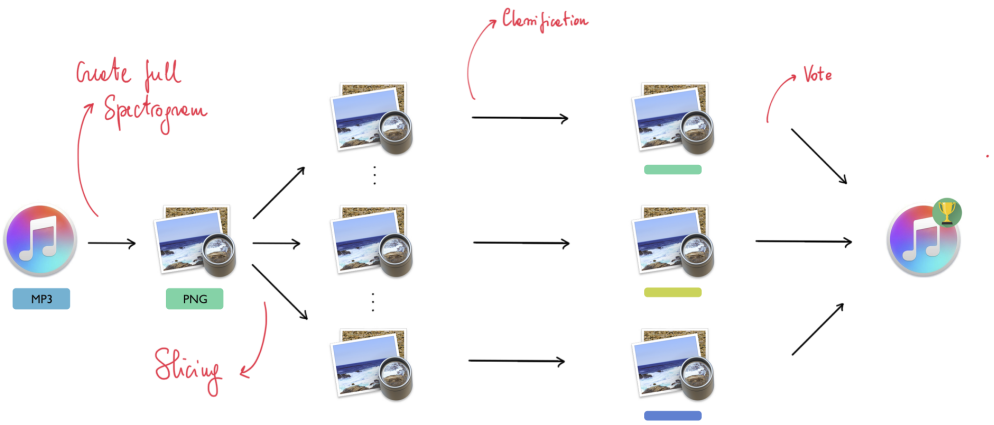

Figure 2. Classification pipeline.