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 [6], 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.* [4] 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 dataset [7] and tested on the CAL500 dataset [8].

2. Data preparation

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.

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

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.

When citing a multi-author paper, you may save space by using “et alia”, shortened to “*et al.*” (not “*et. al.*” as “*et*” is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: “ Frobnication has been trendy lately. It was introduced by Alpher [1], and subsequently developed by Alpher and Fotheringham-Smythe [2], and Alpher *et al.* [3].”

This is incorrect: “... subsequently developed by Alpher *et al.* [2] ...” because reference [2] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al.*

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [2, 1, 5] to [1, 2, 5].

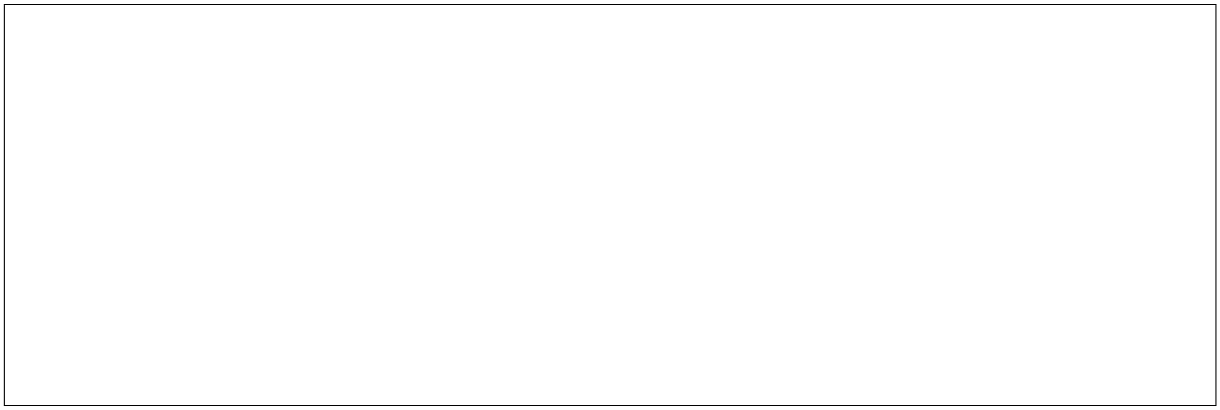

Figure 2. Example of a short caption, which should be centered.

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

Table 1. Results. Ours is better.

6. Data preparation

6.1. 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.

6.2. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [5]. Where appropriate, include the name(s) of editors of referenced books.

6.3. 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}
```

References

[1] FirstName Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002.

[2] FirstName Alpher and FirstName Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003.

[3] FirstName Alpher, FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004.

[4] 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.

[5] Authors. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`.

[6] Piotr Kozakowski and Bartosz Michalak. Music genre recognition. 2016.

[7] 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.

[8] 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.

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