TAMING WILD HORSES WITH ESSENTIA MUSIC EXTRACTOR

Dmitry Bogdanov, Alastair Porter, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
dmitry.bogdanov,alastair.porter,xavier.serra@upf.edu

ABSTRACT

We overview the work done on automatic music audio description using the Essentia Music Extractor. We have successfully applied this tool to the analysis on a large scale in the AcousticBrainz project, in which we have been able to annotate over 3 million music recordings. We are challenged to constantly improve feature extraction algorithms and make our tools more useful for researchers. We reflect on the quality of our current data and identify how it can be improved in following versions to be in line with the essential goals of Music Information Retrieval of building final usable systems. We call on the community for help and contribution.

1. INTRODUCTION

Current Music Information Retrieval (MIR) systems allow for the computation of a multitude of music descriptors based on the audio signal of music pieces. Many of these descriptors are low-level, representing some objective characteristics of the audio that may be impractical and irrelevant for human interpretation. Still, these descriptors are used for employing machine learning algorithms in order to infer more meaningful human-like annotations, but the resulting systems are often inaccurate for practical applications. Some algorithms may work with irrelevant characteristics and therefore these systems are "horses", learning something irrelevant to human understanding of music [3]. Having encountered a frustrating bottleneck in performance of MIR systems, some researchers highlighted the importance of achieving a deeper understanding of the very essence of music by future systems, and called to focus research on music description models that are in line with music perception by human listeners instead of blunt application of signal processing algorithms for data annotation [5, 6].

2. THE ESSENTIA MUSIC EXTRACTOR

The Essentia Music Extractor is an open-source music feature extractor released under Affero GPLv3 license. It is based on algorithms in the Essentia [1] audio analysis library which has been used in a number of industrial applications and has received growing attention within MIR community during the past few years [1]. The extractor computes a large set of descriptors and is optimized for fast analysis suitable for batch computations on large music collections. Many of the descriptors are computed on a frame level and are summarized by statistical distribution across frames. The extractor computes:

- *Spectral descriptors*: spectral shape, Bark/Mel/ERB bands, MFCC, spectral energy, flux, dissonance, complexity, etc.
- *Time-domain and rhythmic descriptors*: loudness, dynamics, onsets, beats, BPM, BPM histogram, beats loudness, danceability, etc.
- *Tonal descriptors*: tuning frequency, chroma, key, scale, chords, etc.
- *Semantic ("high-level") descriptors*: genres, moods (happy, sad), instrumentation (acoustic, electronic, voice gender, timbre color), etc.

The first three categories of descriptors are computed directly from the audio signal, while semantic descriptors are inferred from these signal-based descriptors using high-level models (currently Support Vector Machines (SVMs)) trained on datasets [2], incuding some which are commonly used in MIR. More detailed information about the provided features, including references to the employed MIR and audio analysis algorithms, is provided in the official documentation for Essentia, [2] or by reviewing the code. [3] We distribute this extractor, written in C++, through our website [4] as a static binary for Windows, OSX, and Linux.

3. PROBLEMS AND OPEN QUESTIONS

We have employed Music Extractor in a new platform, AcousticBrainz [5] [2], dedicated to assist with the gathering of musical data from the music enthusiast and research community, and to provide researchers with large

[1] http://essentia.upf.edu
[2] http://essentia.upf.edu/documentation/streaming_extractor_music.html
[3] http://github.com/MTG/essentia/tree/master/src/examples
[4] http://acousticbrainz.org/download
[5] http://acousticbrainz.org

datasets of audio features to work with. Currently AcousticBrainz counts over 3 million music recordings analyzed by its users using our feature extractor. By presenting this project to the MIR community we are challenged to constantly improve feature extraction algorithms and make our tools more useful for researchers. We are still far from building accurate systems and propose a number of open questions. We are asking the community for feedback and help in identifying and implementing missing functionality be it better signal processing algorithms, novel music descriptors, better machine learning models and training datasets, etc.

Genre, mood, and other music classification tasks are examples of problems showing how efficient our tool is in learning concepts relevant to human listeners, the final users of MIR systems. Table 3 summarizes our current SVM model accuracies. It includes values obtained in cross-validation while training the models, as well as accuracy estimations on large holdout data for a number of genre models. While cross-validation accuracies might seem relatively high, our informal experiments using a dump of data gathered on AcousticBrainz have shown that our genre models do not generalize well when applied on large music collections. We doubt about usefulness of our models in practical applications where we would expect the accuracy to be close to 100%. It is unknown to us whether other models generalize well, but we can hypothesize the opposite untill proper evaluation.

We observed unsatisfactory performance of our models on the example of genre classification, and are challenged to tame those wild "horses" and turn them into efficient annotation algorithms. To this end, we struggle to avoid the "garbage in, garbage out" situation and therefore will need to develop framework for estimation of descriptors' accuracy and their robustness in respect to different audio encodings [4]. Furthermore we aim to add more and better descriptors to the Music Extractor and we need to identify those relevant for semantic annotation tasks.

Importantly, we are missing more descriptors that are *musically meaningful* as opposed to simple frame-based acoustic features [5, 6]. For example, *temporal modeling* and the *structure of music* is an important aspect missing in our system. Music Extractor is able to provide frame data, but in the context of AcousticBrainz project frame-wise descriptors are summarized by their statistical distribution over the whole music piece [6] in order to reduce data size and to avoid possible legal restrictions. Our high-level models work with such summarizations and the temporal evolution of many descriptors is lost. However, bag-of-frames approach is inadequate to many music classification tasks and more temporal descriptors that are perceptually relevant for music listeners are required [5].

Apart from improving music descriptors, we identify the need to build better ground-truth datasets. We question what are the requisites for such datasets and envision that they can be built in a semi-automated way or by crowd-sourcing manual annotations. AcousticBrainz provides tools to create such annotations and build classifier models [2]. We anticipate the participation of the MIR community in creation of better datasets using these tools.

[6] We are questioning alternative ways of frame data summarization, such providing statistics over music segments.

Name	Accuracy
Genre Dortmund	$59\% \rightarrow 28\%$
Genre Tzanetakis	$76\% \rightarrow 8\%$
Genre Rosamerica	$87\% \rightarrow 48\%$
Genre electronic	92%
Genre Lastfm	$57\% \rightarrow 58\%$
Mood acoustic	92%
Mood electronic	85%
Mood happy	83%
Mood sad	87%
Mood aggressive	97%
Mood relaxed	93%
Mood party	88%
Moods mirex	58%
Danceability	92%
Voice/instrumental	94%
Dark/bright timbre	94%
Tonal/atonal	98%
Mirex ballroom	70%
Gender	86%

Table 1: Classification accuracies for SVM models currently available in Music Extractor. The provided values are estimated in cross-validation while training the models and on a large holdout set of $\approx 260,000$ annotated music tracks in the case of several genre collections (such accuracies are marked after the $\rightarrow$ symbol).

4. REFERENCES

[1] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J.R. Zapata, and X. Serra. Essentia: An audio analysis library for music information retrieval. In *International Society for Music Information Retrieval (ISMIR'13) Conference*, pages 493–498, 2013.

[2] Alastair Porter, Dmitry Bogdanov, Robert Kaye, Roman Tsukanov, and Xavier Serra. Acousticbrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval (ISMIR'15) Conference*, 2015.

[3] Bob L Sturm. A simple method to determine if a music information retrieval system is a horse. *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

[4] J. Urbano, D. Bogdanov, P. Herrera, E. Gómez, and X. Serra. What is the effect of audio quality on the robustness of mfccs and chroma features? In *International Society for Music Information Retrieval (ISMIR'14) Conference*, pages 573–8, 2014.

[5] Gerhard Widmer. Getting closer to the essence of music: The con espressione manifesto. *ACM Transactions on Intelligent Systems and Technology*, In Press.

[6] Geraint Wiggins et al. Semantic gap?? schemantic schmap!! methodological considerations in the scientific study of music. In *IEEE International Symposium on Multimedia (ISM'09)*, pages 477–482. IEEE, 2009.