MusiClef 2013: Soundtrack Selection for Commercials

Cynthia C. S. Liem
Delft University of Technology
C.C.S.Liem@tudelft.nl

Nicola Orio
University of Padua
Nicola.Orio@unipd.it

Geoffroy Peeters
UMR STMS IRCAM-CNRS
Geoffroy.Peeters@ircam.fr

Markus Schedl
Johannes Kepler University
Markus.Schedl@jku.at

ABSTRACT

MusiClef was one of the “brave new tasks” at MediaEval 2013 with a multimodal approach that combined music, video and textual information in order to evaluate systems that recommend a music soundtrack given the video of a commercial and the information on the product to be advertised.

1. INTRODUCTION

The MusiClef 2013: “Soundtrack Selection for Commercials” task aims at analyzing music usage in TV commercials and determining music that fits a given commercial video. Usually, this task is carried out by music consultants, who select a song to advertise a particular brand or a given product. The MusiClef benchmarking activity, in contrast, aims at making this process automated by taking into account both context- and content-based information about the video, the brand, and the music. The goal of MusiClef 2013, which as in its tradition is motivated by a real professional application, can be summarized as follows: *Given a TV commercial, predict the most suitable music from a list of candidate songs.*

The selection of a suitable soundtrack for a given commercial can be based on a number of characteristics, which have been taken into account while organizing this brave new task. On the one hand, each brand/product has a particular signature that should be underlined by the soundtrack. For this reason, a number of web pages describing either the brands or the products included in the evaluation campaign have been crawled automatically to extract a number of contextual descriptors. One the other hand, the choice of a particular song depends also on the public image of the performer. Again, web pages describing the artists included in the evaluation campaign have been automatically crawled to extract additional contextual descriptors regarding music. Finally, the choice of a soundtrack depends also on how previous commercials were perceived by the public. Thus, as an additional semantic data source, we provide the comments on the commercials made by the persons who uploaded the videos on the web.

Content plays an equally important role in the selection of soundtracks. For this reason a number of descriptors were computed from the audiovisual content of the commercial videos. Clearly, the soundtrack of a commercial video may contain also speech and environmental sounds that are usually not available to music consultants at the time of soundtrack selection. In order to better simulate the selection, we computed the same set of audio descriptors also from the original recordings. It is important to note that, for obvious copyright reasons, we did not distribute the original content but only the lossy descriptors. Participants were referred to web services run by third parties to access to the original multimedia content, both for videos and for songs.

This has been a challenging task, in which multimodal information sources needed to be considered, which do not trivially connect to each other. In particular, participants were asked to provided at least one run based on the combination of multimodal information.

2. THE DATASETS

Two datasets have been made available to participants. First of all, the *development set* included YouTube links to 392 commercial videos for which music has been identified. For each video the development set contained *metadata* on the commercial as available from comments in the YouTube page, *video features* (MPEG-7 Motion Activity and Scalable Color Descriptor [2]), *web pages* about the respective brands and music artists, and *music features* (the well-known MFCC, BLF as proposed in [4], PS209 as proposed in [3], and beat, key, harmonic pattern using the software available at [1]) computed from both the original soundtracks and from the corresponding recordings. Moreover, a set of 227 additional commercial videos has been included in the development set although it was not possibile to identify the original soundtracks. For these videos the same information has been made available, except for music features of the original recordings.

The *test set* included 55 additional commercial videos for which participants have to suggest a suitable soundtrack from 5000 candidate recordings of published music made available from a broadcasting company database. Particular care has been paid to not include the original recording of the commercial in the list of candidate songs. Moreover, the 5000 candidate songs were recorded by the same pool of artists of the development set. To prevent the task becoming a simple audio comparison task, test set videos were provided in muted form. Therefore, for test set videos, no original soundtrack features were provided. However, for the rest, the same information was made available as with videos from the development set. As for the 5000 audio candidate recordings, for each recording a 30 second snippet was extracted, for which the same music features as in the development set were computed.

MediaEval Workshop, 2013. Barcelona, Spain

Audio similarity has been precomputed by the organizers and made available to participants for both sets, in order to provide a common background for all experiments. Participants were free to carry out further processing both on the audio/video features and on the computation of similarity.

3. COLLECTING THE DATA

The process of collecting the data described in the previous sections required a number of steps, that have been carried out by the organizers. In the following we summarize the procedure in order to highlight the main points and to discuss the main decisions that have been taken.

First of all, we selected a number of representative commercials that were available for download on YouTube. We started from a list of annotated commercials proposed in `http://admusicdb.com/`. Starting from this list we automatically crawled YouTube in order to get the complete videos (for this content type, only derived features are distributed), the description inserted by the uploader, and the comments by other viewers.

The audio tracks of the commercials were analyzed by a software for audio fingerprinting and matched with a reference collection of about 380,000 commercial MP3s, which was available thanks to a collaboration with the Italian broadcaster RTI. Only about 50% of the original soundtracks were successfully identified, thus we manually inspected the reasons for missing identifications. In general, a number of soundtracks were composed purposely for the commercials while some of them were played live by the testimonials. The remaining soundtracks were simply not present in the reference collection or were stored as different covers in the reference collection. In order to deal with the latter, we collected all the available covers and manually compared their music content with the soundtracks, evaluating the similarity in a three-level scale. Through manual identification we increased the available MP3s to about 60% of the downloaded videos. Participants were informed, for each MP3, on the confidence level of the identification.

The final step consisted in selecting the files for the actual task: videos and MP3s. Videos were chosen among the ones where no identification was possible, selecting the ones with a similar length of about 30 seconds. MP3s were selected as a subset of the reference collection, taking particular care that they were performed by the same pool of artists and that did not contain the original songs. For each MP3 we extracted a sample of 30 seconds that were used for the task.

In parallel with the content descriptors, we retrieve relevant contextual information. Starting from the complete list of videos, we could select the set of brands and products that have been advertised and the set of artists that were mentioned as the main performers. We crawled the eb submitting three different queries to Bing search engine: "*brand/product*", "*artist* music", and "*artist* music review". In order to guarantee reproducibility of the results, we downloaded the complete pages besides computing the Lucene index and the term weight (TF × IDF).

4. EVALUATION

Participants could submit one to three runs, with the requirement that at least one run should use multimodal information. For each video in the test set, participants are requested to propose a ranked list of 5 candidate songs.

Evaluation is carried out using the Amazon Mechanical Turk platform. For every video in the test set, a HIT was designed presenting the muted test set video, and all top-5 song (snippet) suggestions for the video, as submitted by the participants. These song suggestions were presented in randomized order. For each HIT, 5 assignments are released. Since both the video and each of the song snippets were not longer than 30 seconds, the load on the side of the workers was kept within reasonable bounds.

MTurk workers are asked to grade the suitability of each song suggestion on a 4-level Likert scale, ranging from *very poor* (1 point) to *very good* (4 points). There also is a fallback 'impossible to tell' option, which required a mandatory explanation on why the suitability could not be graded.

For each run, evaluation results are computed using three different measures. Let V be the full collection of test set videos, and let $\overline{s_r(v)}$ be the average suitability score for the audio file suggested at rank r for video v. Then, the evaluation measures are computed as follows:

- Average suitability score of the first-ranked song:
$$\frac{1}{|V|}\sum_{i=1}^{|V|}\overline{s_1(v_i)}$$
- Average suitability score of the full top-5:
$$\frac{1}{|V|}\sum_{i=1}^{|V|}\frac{1}{5}\sum_{r=1}^{5}\overline{s_r(v_i)}$$
- Weighted average suitability score of the full top-5. Here, we apply a weighted harmonic mean score instead of an arithmetic mean:
$$\frac{1}{|V|}\sum_{i=1}^{|V|}\frac{\sum_{r=1}^{5}\overline{s_r(v_i)}}{\sum_{r=1}^{5}\frac{\overline{s_r(v_i)}}{r}}$$

It should be stressed that this brave new task is highly novel and non-trivial in terms of 'ground truth'. This is why we purely use human ratings for the evaluation, and use the different measures above to both study rating and ranking aspects of the results.

Acknowledgments

MusiClef has been partially supported by EU-FP7 through the PHENICX project (no. 601166) and the PROMISE Network of Excellence (no. 258191); it is also partially supported by the Austrian Science Funds (FWF): PP22856-N23 and P25655. The work of Cynthia Liem is supported in part by the Google European Doctoral Fellowship in Multimedia.

5. REFERENCES

[1] Ircam. Analyse-Synthèse: Software. `http://anasynth.ircam.fr/home/software`. Accessed: Sept. 2013.

[2] B. S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface.* John Wiley & Sons, New York, USA, 2002.

[3] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On Rhythm and General Music Similarity. In *Proc. of ISMIR*, 2009.

[4] K. Seyerlehner, G. Widmer, and T. Pohle. Fusing Block-Level Features for Music Similarity Estimation. In *Proc. of DAFx*, 2010.