MediaEval 2014: A Multimodal Approach to Drop Detection in Electronic Dance Music *

Anna Aljanaki
Information and Computing Sciences
Utrecht University
the Netherlands
a.aljanaki@uu.nl

Mohammad Soleymani
Computer Science Dept.
University of Geneva
Switzerland
mohammad.soleymani@unige.ch

Frans Wiering
Information and Computing Sciences
Utrecht University
the Netherlands
F.Wiering@uu.nl

Remco C. Veltkamp
Information and Computing Sciences
Utrecht University
the Netherlands
R.C.Veltkamp@uu.nl

ABSTRACT

We predict drops in electronic dance music (EDM), employing different multimodal approaches. We combine three sources of data: noisy labels collected through crowdsourcing, timed comments from SoundCloud and audio content analysis. We predict the correct labels from the noisy labels using the majority vote and Dawid-Skene methods. We also employ timed comments from SoundCloud users to count the occurrence of specific terms near the potential drop event, and, finally, we conduct an acoustic analysis of the audio excerpts. The best results are obtained, when both annotations, metadata and audio, are combined, though the differences between them are not significant.

1. INTRODUCTION

This working notes paper describes a submission to the CrowdSorting brave new task in the MultiMediaeval 2014 benchmark. The main aim of the task is to detect drops in electronic music. According to the Wikipedia definition: "Drop or climax is the point in a music track where a switch of rhythm or bassline occurs and usually follows a recognizable build section and break"[1]. The task involves categorizing 15 second electronic music excerpts into three categories: those containing a drop, those containing part of the drop, and those without a drop. The organizers provide three types of data: unreliable crowdsourced annotations, timed comments from SoundCloud users, and audio. Acoustic analysis is optional to the task. For more detail we refer to the task overview paper [3].

We submitted four runs: three are based on annotations and other metadata, and one is based on a combination of metadata and acoustic features. Due to the social attention that drop phenomenon gets in electronic music, the task of drop detection is naturally suitable for a combined approach, using both metadata and acoustic features. The acoustic-only approach is rather challenging, because there are many informal descriptions of what constitutes a drop, including rhythmic and dynamical changes, or specific patterns in the bass line. Also, the presence or absence of drop in a specific case is debatable.

*First two authors contributed equally to this work and appear in alphabetical order.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

2. RELATED WORK

Karthik Yadati et al. [4] (the organisers of Mediaeval 2014 CrowdSorting task) conducted an acoustic analysis to detect drops in EDM. The audio was first segmented under the assumption that a drop moment must be an important structural boundary. Then, each of the segmentation boundaries was classified based on the analysis of several features in a time window around the potential drop. MFCCs, spectrogram and rhythmical features were used based on the notion that a drop event is usually characterized by a sudden change of rhythm and timbre.

3. APPROACH

For each of the excerpts, three annotations from MTurk workers were provided. Fleiss' kappa for these labels was 0.24 (calculated without songs from the fourth category, "absent sound file"). Around 30% of the excerpts were unanimously rated by annotators. For about 60%, two of the annotators agreed. For the remaining 10% of the excerpts, all the annotators provided different answers. We mainly sought to improve the categorization of the second and especially the last categories.

3.1 Metadata analysis and improving ground truth

The first run employs a simple majority vote. In case all the annotators categorize the segment differently, we label it as containing part of the drop.

In the second run, we use the Dawid-Skene algorithm [2] to compute the probabilities of each label, and the quality of workers, based on their agreement with other workers. The Dawid-Skene model calculates the confusion matrices for each worker using a Maximum-Likelihood estimation based on their agreement with the other workers. We use the Get-another-Label toolbox[1] implementation of Dawid-Skene. Then, we use the calculated probabilities combined with the given labels to predict the actual labels.

[1]https://github.com/ipeirotis/Get-Another-Label

In the third run, we count the number of timed comments from SoundCloud users which include the term ”drop” near the moment of hypothetic drop (the 15 second time window defined by organizers). We use a Naïve Bayes classifier to train a model based on a number of comments in addition to the three noisy labels. The model is only used to categorize the excerpts with no agreement between annotators.

3.2 Audio analysis

As a training data, we employed the excerpts for which all the three workers agreed. There were 164 such excerpts in total, 105 for which workers indicated that the excerpt contained an entire drop, 54 for which they indicated there was no drop, and 4 for which they agreed there was part of the drop present. We decided to exclude the excerpts labeled ”part of the drop”, as it is not possible to learn to recognize it based on just four samples.

The acoustic approach was based on the fact that during a drop, there is usually a moment of silence, or sometimes the loudness level changes drastically after the drop. We analyzed the energy of the signal in non-overlapping windows of 100 ms. The obtained time-series was smoothed using the weighted moving average. The smoothed time-series was segmented on their local maximums and minimums. To predict the presence of the drop event, we used the following statistics on these events:

1. The value of the biggest local minimum in an excerpt
2. The fraction of the biggest minimum to an average minimum
3. The number of potential drop events, as detected by decrease in loudness bigger than threshold
4. The dynamic range of the excerpt

Based on these characteristics and a ground-truth of 160 excerpts, we trained a logistic regression classifier to predict the presence of drops, and obtained 80% precision with 10-fold cross validation. The model was used to predict the presence of drops for the excerpts where all three workers gave different ratings (i.e., ”drop is present”, ”part of the drop is present”, ”drop is not present”). The biggest limitation of this approach is that the model does not incorporate the ”part of the drop” category.

4. EVALUATION

The evaluation metric for this task is the F1 score, calculated based on high-fidelity labels from the experts, used as a ground-truth. Though there are some differences between submissions, none of them were statistically significant on a one-sided Wilcoxon ranksum test. The majority vote scores are as usual hard to beat. Using comments from SoundCloud users results in some improvement, and using acoustic features performs similarly. Looking at the accuracy per category, we can see that the acoustic submission suffers from imprecision in the category ”part of the drop”, which is natural, because it does not model that. On the other hand, the precision of ”no drop” labels is higher than for all other submissions.

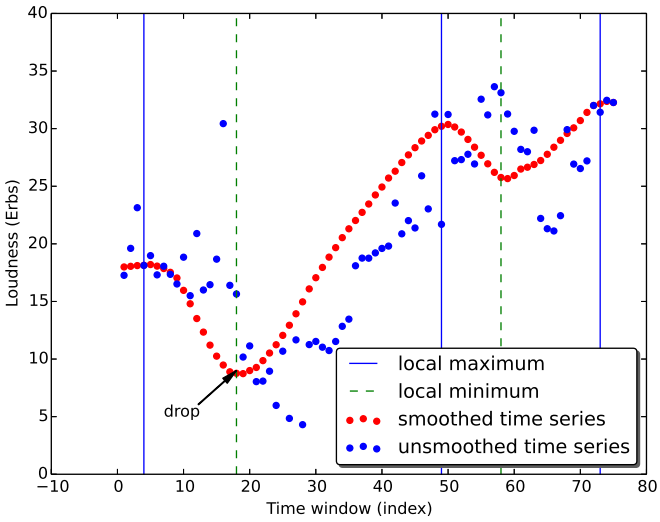

Figure 1: A smoothed and segmented time-series of an excerpt with drop.

Run	Name	F1	Drop	Part	No drop
Run 1	Majority Vote	0.69	0.72	0.31	0.75
Run 2	DS	0.69	0.72	0.31	0.75
Run 3	MV+SoundCloud	0.7	0.73	0.28	0.76
Run 4	MV+Audio	0.71	0.72	0.27	0.79

5. CONCLUSION

In this task, we only achieved marginal improvement over the baseline, i.e., majority vote. Both acoustic analysis and the use of SoundCloud metadata resulted in a small but insignificant prediction improvement. This shows that in the presence of enough labels given by MTurk workers, we could not significantly improve the accuracy based on the content or social media metadata. However, they are nevertheless useful in cold start scenarios.

6. ACKNOWLEDGEMENTS

This publication was supported by the Dutch national program COMMIT.

7. REFERENCES

[1] M. J. Butler. *Unlocking the Groove. Rhythm, Meter, and Musical Design in Electronic Dance Music.* 2006.

[2] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, (1):20–28, 1979.

[3] M. L. Karthik Yadati, Pavala S.N. Chandrasekaran Ayyanathan. Crowdsorting timed comments about music: Foundations for a new crowdsourcing task. In *MediaEval Workshop*, Barcelona, Spain, October 16-17 2014.

[4] K. Yadati, M. A. Larson, C. C. Liem, and A. Hanjalic. Detecting drops in electronic dance music: Content-based approaches to a socially significant music event. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, 2014.