Ref. Ares(2017)3824184 - 31/07/2017

Deliverable D4.5

Evaluation report on the first prototype tool for the automatic semantic description of music pieces

Grant agreement nr	688382
Project full title	Audio Commons: An Ecosystem for Creative Reuse of Audio Content
Project acronym	AudioCommons
Project duration	36 Months (February 2016 - January 2019)
Work package	WP4
Due date	31 July 2017 (M18)
Submission date	31 July 2017 (M18)
Report availability	Public (X), Confidential ( )
Deliverable type	Report (X), Demonstrator ( ), Other ( )
Task leader	QMUL
Authors	Johan Pauwels.
Document status	Draft (), Final (X)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Executive Summary

As part of the Audio Commons Ecosystem, a number of tools are provided for the automatic analysis of audio content without the need for human intervention. These tools are designed for extracting i) musical audio properties for music pieces and music samples, and ii) non-musical audio properties for any kind of sounds. Work-in-progress versions of the tools extracting audio properties for music samples and music pieces have been released in conjunction with deliverables D4.2 and D4.3.

In this document, we reflect on the tools for the automatic semantic description of music pieces presented in D4.3. The aim is to inform future work on this topic, which will culminate in a second release of tools for the semantic annotation of music pieces (in D4.8). It is the counterpart for music pieces of D4.4, which evaluates the tools for annotation of music samples presented in D4.2.

Since the annotation of music pieces as a whole is a much more popular research subject than annotating music samples, the performance of labelling algorithms is well known from the literature. A large contributing factor is the yearly recurring Music Information Retrieval Evaluation eXchange (MIREX), in which a large number of labelling tasks get evaluated on a variety of datasets. Because we don't have any new datasets with music piece annotations available compared to the ones that are used for MIREX, we didn't rerun those experiments for this deliverable, but refer to the MIREX website for detailed results. Some general conclusion we can draw from those results is that most labelling task are mature in the sense that no big improvements in the state-of-the-art performance are to be expected in the near future. At the same time, that performance shows a strong dependency on the type of data and doesn't attain a level comparable to human annotations in most cases. For the deployment of those algorithms in the Audio Commons Ecosystem, with its wide variety of audio data, their raw performance is simply not enough to provide the content user with a satisfying experience. Therefore we focus in this deliverable on ways to increase user experience by means of confidence measures.

Similar to previous work for tempo estimation on music loops (see D4.2), we present confidence measures that can be used to filter Audio Commons content for musical properties such that only files for which the quality of annotations is sufficiently high get returned. Contrary to the confidence measures for loop tempo, no specific information about the domain is used, but it is specific to systems that use hidden Markov models. Therefore, the proposed measures are more general in the sense that they can be used for multiple labelling tasks, but less general in the sense that they are tied to one specific approach to accomplish a certain task. We illustrate the applicability of the new confidence measures with a chord estimation task. We investigate both the capabilities of the confidence measure to select files with high-quality chord sequences and to select the optimal channel for chord estimation in a multi-track scenario.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Background

Inspired by previous work in WP4 on confidence measures for tempo estimation in loops [Font16], we explore ways to add confidence measures to other music labelling tasks. It is useful to start by reflecting on the underlying conditions that made Font and Serra's work on loops successful. They managed to define a confidence measure that can be calculated from just the generated BPM value and the file duration. It is therefore oblivious to the method used to calculate the BPM. The advantage is that knowledge of and access to the inner workings of the algorithm are not required in order to use the confidence measure. However, this type of confidence measure relies on extra prior knowledge about the application domain, which is used to verify the output against. This domain knowledge is the only information we have in the absence of internal states of the algorithm and (obviously) the target labels. Therefore we call confidence measures that do not rely on the internal state of an algorithm, but exclusively on knowledge about its application domain *domain-specific*, as opposed to *algorithm-specific* measures. For example, the domain knowledge used for loops is that they are cut in such a way that each loop contains an exact number of beats.

It is of the utmost importance that this domain information hasn't been exploited yet by the estimation algorithm. All tested software in D4.2 and [Font16] fulfils this criterion, as they were developed for music pieces in general, not just loops. If algorithms would already rely on this prior knowledge, the output would be internally adjusted by, for example, keeping only those tempo candidates that lead to an exact multiple of beats for the duration of the loop. The confidence measure would then always be maximal and therefore useless.

Finding such unexploited knowledge for a specific application is not always possible, and if there is one, it also needs to be distinctive enough. Take for instance the case of key estimation in loops. A reasonable prior would be to assume that there are no key changes for the duration of a loop. Even if a key estimation algorithm is capable of producing key changes, it is unlikely that multiple keys will be produced over the short duration of a typical loop. No reliable confidence measure can then be derived from this additional information.

We argue that having sufficient unexploited and distinctive domain knowledge for a particular task is a rare event. In practice, it is therefore more likely that we need recourse to algorithm-specific confidence measures. These are defined using the intermediate states of the algorithm, which unfortunately means that separate measures need to be formulated for each algorithm and that the resulting confidence cannot easily be compared between algorithms. The upside is that they are not tied to a particular domain.

In the remainder of this document, we propose some candidate algorithm-based confidence measures. To mitigate their algorithm-specificity, we will look at the framework of hidden Markov Models [Rabiner1989] which is commonly used in a variety of estimation tasks. Our hope is that the proposed solutions will therefore be task-independent and widely applicable.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

1 Introduction

Most of the efforts in music information retrieval research are directed towards improving the performance of various automatic labelling tasks. This consists of developing algorithms that are increasingly good at approximating some reference labels, often produced by human annotators, based on an input audio file. These labels represent different musical concepts, such as genre, tempo, instrumentation or musical key. In the Audio Commons Ecosystem, the ontology defined in WP2 determines the labels that are applicable for each content creator.

When such algorithms are deployed as tools for automatic describing the content in the Audio Commons Ecosystem, however, no explicit comparison is made between the generated labels and a reference. An example is the retrieval of audio based on musically meaningful search terms. The only relevant measure of performance here is the degree of satisfaction of the content user with the returned audio files. The user will subconsciously verify if the returned audio corresponds somewhat to the query term, and be dissatisfied if it doesn't, but this implicit and informal evaluation is nowhere as rigorous as the numerical evaluation performed to demonstrate algorithmic improvements. This gap between algorithmic evaluation and user evaluation makes that increases in algorithmic performance do not necessary lead to increases in user satisfaction.

Crucially, in many retrieval tasks (e.g. searching on Freesound or Jamendo) the precision is more important than the recall. The users only judge the relevance of the returned files, the amount of potentially useful audio files that are not returned to them are unknown and irrelevant (once the amount of returned files reaches a minimally acceptable number of course). A relatively easy way to improve the perceived quality of the selection (and thereby user satisfaction) would be to only return those files for which the generated labels are known to be of a high quality. This necessitates a reliable measure of confidence for the generated output labels, which must be calculated without relying on a known reference output.

1.1 Main objectives and goals

Despite its obvious use-case, not much work has been performed on confidence measures for music labelling. For tempo estimation in music loops specifically, a thorough study has already been performed by Font and Serra [Font16]. They propose a new confidence measure and compare it to earlier efforts of Zapata et al. [Zapata12]. In this document, we evaluate new methods for confidence estimation for a chord estimation system based on a hidden Markov model (HMM). We test their application for filtering low-quality output from datasets in Section 2.1. Next, a novel application for confidence measures is discussed in Section 2.2. We end by formulating some conclusions and directions for future work in Section 3.

1.2 Terminology

AudioCommons: reference to the EC H2020 funded project AudioCommons, with grant agreement nr 688382.

Audio Commons Initiative: reference to the AudioCommons project core ideas beyond the lifetime and specific scope of the funded project. The term “Audio Commons Initiative” is used to imply i) our will to continue supporting the Audio Commons Ecosystem and its ideas after the lifetime of the funded project, and ii) our will to engage new stakeholders which are not officially part of the project consortium.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 688382

Audio Commons: generic reference to the Audio Commons core ideas, without distinguishing between the concept of the initiative and the actual funded project.

Audio Commons Ecosystem (ACE): set of interconnected tools, technologies, content, users and other actors involved in publishing and consuming Audio Commons content.

Audio Commons content (ACC): audio content released under Creative Commons licenses and enhanced with meaningful contextual information (e.g., annotations, license information) that enables its publication in the ACE.

Content creator: individual users, industries or other actors that create audio content and publish in the ACE through content providers.

Content provider: services that expose content created by content creators to the ACE.

Content user: individual users, industries or other actors that use the content exposed by content providers and created by content creators in their creative workflows.

Tool developer: individual users, industries or other actors that develop tools for consuming (and also potentially publishing) Audio Commons content.

Embeddable tools: tools for consuming Audio Commons content that can be embedded in existing production workflows of creative industries.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

2 Confidence measures for hidden Markov model-based chord estimation

In [Pauwels17], we presented two confidence measures for music labelling systems based on hidden Markov models. The first one is based on the median contribution per frame of the optimal chord sequence returned by the maximum a posteriori (MAP) decoding of the HMM. Its advantage is that it requires almost no computational overhead compared to regular HMM decoding. The second, more computationally intensive one, is called *pointwise path difference (PPD)*. It is derived from the differences in optimal chord sequence as returned by MAP decoding and by piecewise maximum a posteriori (PMAP) decoding.

The experiments in the remainder of this document are performed using a chord estimation system that uses either DeepChroma [Korzeniowski16] or Compressed Log Pitch (CLP) [Mueller11] features. The system consists of a baseline HMM with uniform initialisation probabilities and a transition matrix where all diagonal elements are set to the same constant and the remaining probability is distributed over the off-diagonal elements. This constant was experimentally set to 0.1. The observation probabilities are calculated by taking the cosine similarity with binary templates that are set to 1 whenever a chroma is theoretically present in a chord and 0 otherwise. The hidden search space consists of 48 chords: major, minor, diminished and augmented triads for each of 12 possible roots.

We use two datasets for testing the chord estimation systems and their confidence measures, in order to investigate data-specific behaviour. The first is the “Isophonics” dataset [Mauch09]. Specifically, we use the subset that is used for the MIREX chord estimation task. It contains 217 songs and is comprised of 12 Beatles albums (180 songs), a Queen compilation (19 songs) and one Zweieck album (18 songs). The second is the “RWC Popular” dataset [Goto02]. The latter contains 100 Japanese pop songs purposefully recorded for music information retrieval research.

2.1 Filtering datasets based on confidence

As we have two candidate confidence measures, we can systematically test them on the four combinations of features and datasets. Therefore we perform a similar experiment as in [Font16]. We start by taking the duration-weighted average score over the complete dataset and then progressively filter the files by first excluding those for which the confidence measure is the lowest. A good confidence measure will then lead to a monotonic increase in score as the filtering threshold increases. The results for the Isophonics dataset and the RWC Popular dataset can be found in Figure 1.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 688382

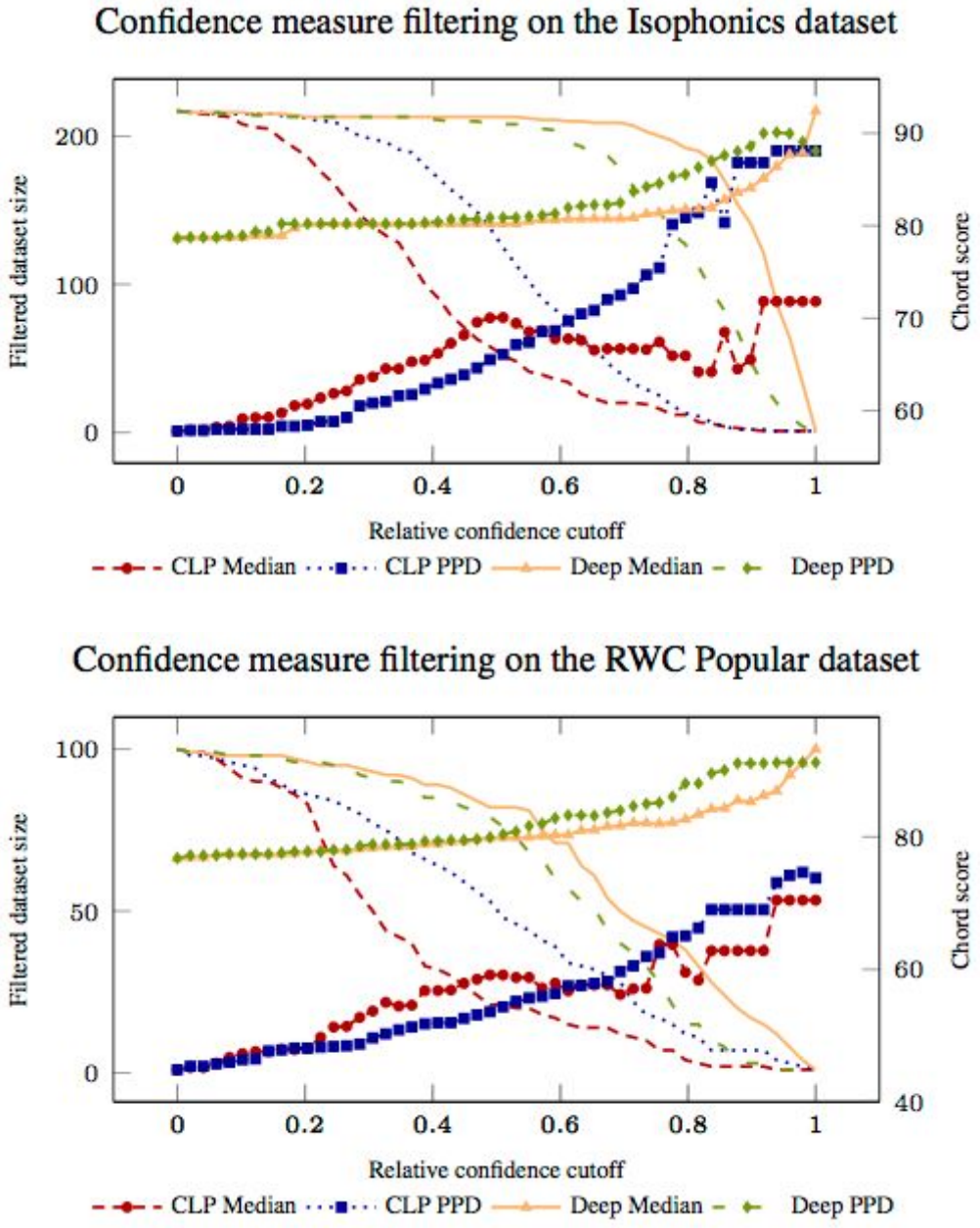

Figure 1: Confidence filtered chord scores for two datasets. The marks indicate the average score over the filtered dataset. The full lines represent the number of remaining files in the database as a function of the confidence cutoff.

In all cases, we observe that the PPD is working well as a confidence measure. The score of the filtered dataset increases monotonically, save for a few exceptions when the number of remaining songs in the dataset becomes so low that the average scores become noisy. The curves of the filtered data size as a function of confidence cutoff are also close to straight, which means the PPD is nearly linearly distributed between its extrema. As expected, the median of the per frame contributions to the optimal path is less suitable as a confidence measure. The filtered score initially increases in all situations though, so it can still be used to remove the files with the lowest confidence from the dataset. Doing so will increase the precision when looking for a particular chord sequence in a dataset, for example, at the expense of decreasing the recall. Particularly for the CLP features, the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

median-based confidence measure seems to work less well. A possible explanation is that the neural network based chromas take context into account. The observations derived from the CLP features are therefore noisier, which will affect the median more.

2.2 Automatic optimal channel selection

In this section, we explore an alternative usage for confidence measures. Traditionally, labels in music information retrieval are estimated from mixed down mono audio. Using the mono mix ensures that all information present in the audio is used for the label estimation. For certain types of labels however, it might be beneficial to selectively ignore some of the information. For example, ignoring percussion while estimating chords can be helpful, which has led to percussion separation techniques [Ueda10].

If we have multi-channel audio at our disposal, it is therefore possible that analysing a specific channel or combination of channels leads to higher quality labels than when the mono mixdown is used. The problem is then how to determine this (combination of) channel(s). Next, we'll verify if a confidence measure can be used for this.

Ideally, we'd use multi-channel or multi-track audio for this experiment, but since there is no such dataset with annotated chords, we propose an alternative. Starting from stereo Isophonics audio files, we demix the left (L) and right (R) channel according to their panning position into centre (C), hard left (HL) and hard right (HR). We employ the technique used by the "center cut" audio filter of the open-source video editor VirtualDub. It operates in the complex spectral domain and relies on the fact that HL and HR are perpendicular to each other, such that L = C + HL and R = C + HR. In addition to these channels, we calculate the mono (L + R) and sides (HL + HR), such that we end up with seven virtual channels per song.

For each channel, we estimate the chord sequences from CLP features and their confidences. We first aim to determine the theoretical limits of optimal channel selection by performing an oracle-style experiment where we select the channels that lead to the biggest increase and biggest decrease in chord score when compared to the reference mono channel. Then we check how well we can retrieve the optimal channel by selecting the one that returns the chord sequence with the highest confidence. Both confidence measures are tested.

The channel selection results overall and per album are reported in Table 1. From the oracle experiments we learn that a sizeable improvement in chord score can potentially be achieved by selecting the optimal channel, but also that the consequences of choosing the wrong channel can be severe. The PPD measure can be used successfully to determine a better channel than the mono reference, and manages to reach a bit more than half of the score improvement attained by the oracle. The median-based confidence measure, on the other hand, is not suitable to select the optimal channel.

Based on the individual results per album, no relation with mixing style can be established. The mixing practices range from the mono-like early Beatles albums to the hard-panned late Beatles albums, with more modern Queen and Zweieck in between, but no trends in the (potential) score increase can be identified. Note that when we repeated the experiments with the DeepChroma chord estimation system, the oracle-based maximal increase was barely over 2%-points, and the PPD increase proportionate. A reason might be that the neural network is trained on mono mixes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Album title	Mono score	Oracle best increase	Oracle worst decrease	Median conf. increase	PPD conf. increase
The Beatles - Please Please Me	52.86	6.01	-13.93	-7.23	1.17
The Beatles - With the Beatles	56.22	0.6	-26.81	-18.72	-2.11
The Beatles - A Hard Day's Night	56.21	3.05	-31.91	-21.88	2.24
The Beatles - Beatles for Sale	63.98	9.33	-3.26	5.91	7.33
The Beatles - Help!	52.54	11.82	-13.23	7.9	10.03
The Beatles - Rubber Soul	59.25	5.44	-18.89	-3	2.77
The Beatles - Revolver	66.06	5.64	-17.63	-0.7	3.71
The Beatles - Sgt. Pepper's Lonely Hearts Club Band	51.33	4.93	-19.81	-4.83	-1.74
The Beatles - Magical Mystery Tour	66.77	3.52	-18.33	-3.1	0.54
The Beatles - The Beatles (CD1)	62.45	6.32	-17.82	1.76	1.68
The Beatles - The Beatles (CD2)	52.05	6.16	-18.91	1.38	1.67
The Beatles - Abbey Road	63.86	8.39	-17.47	4.33	4.11
The Beatles - Let It Be	61.16	11.96	-8.09	9.24	8.21
Queen - Greatest Hits I	47.5	7.86	-3.54	6.53	6.66
Queen - Greatest Hits II	66.16	4.02	-5.49	-1.45	-1.5
Zweieck - Zwielicht	54.73	7.73	-8.1	5.35	6.26
Overall	57.81	6.6	-14.97	-0.39	3.42

Table 1: Channel selection results grouped per album, using CLP features. The reference mono channel score is reported along with the absolute score differences for the best and worst oracle-style and the confidence-based channel selection.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

3 Conclusion

In this deliverable, we investigated confidence measures for HMM-based music labelling systems. They were tested for their ability to filter low-quality output of a chord estimation system. Out of the investigated median optimal chord contribution (MOPC) and pointwise path difference (PPD) the latter seemed to perform best, although at a higher computational cost than the former. Depending on the system configuration, an improvement of 5 to 10 absolute percentage-points can be achieved by by filtering out the half of the dataset with the lowest confidence values. Furthermore, the capacity of the confidence measures to select the most optimal channel to use for chord estimation has been evaluated. On average, the PPD managed to select a better channel than the reference mono channel, but there are some instances where the selection based on confidence actually led to a worse result.

In the future, we hope to verify the applicability of the proposed confidence measures to other labelling tasks. Alternative confidence measures will continue to be researched, both for the tasks presented in this document and new ones. The selection of appropriate cutoff values for the confidence also needs to be investigated. Furthermore, we want to explore whether a confidence measure can be used to select the optimal HMM parametrisation. For instance, the chord self-transition probability is currently set to a dataset-wide optimal value, even though it is clearly related to harmonic rhythm and therefore song-dependent. It might be worth investigating if the best value out of a number of candidates can be selected based on confidence.

In the Audio Commons Ecosystem, these confidence measures will be deployed in the next version of the tool for automatic semantic description of music pieces, which will be described in D4.8. We can also investigate whether our findings are applicable to music pieces as well, by modifying the tools presented in D4.2. Those conclusions will then be presented in D4.7.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

4 References

[Font16] Frederic Font and Xavier Serra. Tempo estimation for music loops and a simple confidence measure. In Proceedings of the 17th Conference of the International Society for Music Information Retrieval (ISMIR), pp. 269-275, New York, USA, 2016.

[Goto02] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical and jazz music databases. In Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR), volume~2, pages 287--288, 2002.

[Korzeniowski16] Filip Korzeniowski and Gerhard Widmer. Feature learning for chord recognition: the deep chroma extractor. In Proceedings of the 17th Conference of the International Society for Music Information Retrieval (ISMIR), 2016.

[Mauch09] Matthias Mauch, Chris Cannam, Matthew Davies, Simon Dixon, Chris Harte, Sefki Kolozali, Dan Tidhar, and Mark Sandler. OMRAS2 metadata project 2009. In Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR), 2009.

[Mueller11] Meinard Müller and Sebastian Ewert. Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), pages 215--220, 2011.

[Pauwels17] Johan Pauwels, Ken O'Hanlon, György Fazekas and Mark B. Sandler. Confidence measures and their applications in music labelling systems based on hidden Markov models. In Proceedings of the 18th Conference of the International Society for Music Information Retrieval (ISMIR), 2017.

[Rabiner89] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In Proceedings of the IEEE, 77(2):257--286, February 1989.

[Ueda10] Yushi Ueda, Yuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5518--5521, March 2010.

[Zapata12] Jose Ricardo Zapata, Matthew E.P. Davies, Andre Holzapfel, Joao L Oliveira, and Fabien Gouyon. Assigning a confidence threshold on automatic beat annotation in large datasets. In Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR), pages 157--162, Porto, Portugal, 2012.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382