A case for reproducibility in MIR.
Replication of ‘a highly robust audio fingerprinting system’

Joren Six, Federica Bressan, Marc Leman
IPEM,
Department of Musicology, Ghent University
Ghent, Belgium
joren.six@ugent.be

Abstract

Claims made in many MIR publications are hard to verify due to the fact that (i) often only a textual description is made available and code remains unpublished – leaving many implementation issues uncovered; (ii) copyrights on music limit the sharing of datasets; and (iii) incentives to put effort into reproducible research – publishing and documenting code and specifics on data – is lacking.
In this article the problems around reproducibility are illustrated by replicating a MIR work. The system and evaluation described in ‘A Highly Robust Audio Fingerprinting System’ is replicated as closely as possible. The replication is done with several goals in mind: to describe difficulties in replicating the work and subsequently reflect on guidelines around reproducible research. Added contributions are the verification of the reported work, a publicly available implementation and an evaluation method that is reproducible.

Keywords: Replication, Acoustic fingerprinting, Reproducibility.

1 Introduction

Reproducibility is one of the corner-stones of scientific methodology. A claim made in a scientific publication should be verifiable and the described method should provide enough detail to allow replication, ‘reinforcing the transparency and accountability of research processes’ (Levin et al., 2016, p.129). The Open Science movement has recently gained momentum among publishers, funders, institutions and practicing scientists across all areas of research. It is based on the assumption that promoting ‘openness’ will foster equality, widen participation, and increase productivity and innovation in science. *Re-usability* is a keyword in this context: data must be ‘useful rather than simply available’ (Levin et al., 2016, p.133), with a focus on facilitating the advancement of knowledge based on previous work (spot check to avoid repeating work) rather than on verifying the correctness of previous work.
From a technical standpoint, sharing tools and data has never been easier. Reproducibility, however, remains a problem. Especially for Music Information Retrieval (MIR) research and research involving complex software systems. This problem has several causes:

- Journal articles and especially conference papers have limited space for detailed descriptions of methods or algorithms. Even for only moderately complex systems there are implementation issues which are glossed over in textual descriptions. This makes articles readable and the basic method intelligible, but those issues need to be expounded somewhere. Preferably in documented, runnable code. Unfortunately, **intel-**

lectual property rights by universities or research institutions often limit researchers to distribute their code. This is problematic since it leaves the ones reproducing the work guessing for details and makes replicating a study prohibitively hard.

- **Copyrights on music** make it hard to share music freely. MIR research often has commercial goals and focuses on providing access to commercial, popular music. It is sensible to use commercial music while doing research as well. Unfortunately, this makes it potentially very expensive to reproduce an experiment: all music needs to be purchased again and again by researchers reproducing the work.

  The original research also needs to uniquely identify the music used, which is challenging if there are several versions, re-issues or recordings of a similarly titled track. Audio fingerprinting techniques allow us to share unique identifiers[1] but in practice this is rarely done. When sharing data toghether with annotations, it is best to adhere to the list of best practices given by Peeters and Fort (2012).

- The evaluation of research work (and most importantly of researchers) is currently based on the number of articles published in ranked scientific journals or conferences. Other types of scientific products are not valued as much. The advantage of investing resources in documenting, maintaining and publishing reproducible research and supplementary material is not often obvious in the effort of prioritising and strategising research outputs (Levin et al., 2016, p.130). Short-lived project funding is also a factor that directs attention of researchers to short-term output (publications), and not to long-term aspects of reproducible contributions to a field. In short there is **no incentive** to spend much time on non-textual output.

Reproducing works is not an explicit tradition in computer science research. In the boundless hunt to further the state-of-the-art there seems no time or place for a sudden standstill and reflection on previous work. Implicitly, however, there seems to be a lot of replication going on. It is standard practice to compare results of a new method with earlier methods (baselines) but often it is not clear whether authors reimplemented those baselines or managed to find or modify an implementation. Due to a lack of standardized datasets, approaches are often hard to compare directly. Reimplementation is not only resource consuming, but never gives the guarantee that the re-created code matches the antecedent down to the last detail (Peng, 2011; Mesirov, 2010). Moreover, if a replication and evaluation is done thoroughly this does not yield new findings (if all goes well). It, therefore, may be unlikely to get published. A considerable amount of work that risks remaining unpublished is not a proposition many researchers are looking forward to, expressing the tension for researchers to 'act for the good of the community or their own' (Nosek et al., 2015).

In the social sciences the reproducibility project illustrated that the *results* of many studies could not be successfully reproduced (Pashler and Wagenmakers, 2012; Collaboration et al., 2015) mainly due to small sample sizes and selection bias, a finding that was also demonstrated in a special issue in Musicae Scientiae on *Replication in music psychology* (Fischinger, 2013). In these replicated studies the main problem did not lay in replicating methods.

For research on complex software systems (MIR) it is expected that the replicated results will closely match the original if the method can be accurately replicated and if the data are accessible. But those two conditions are hard to meet. The replication problem lies exactly in the difficulties in *replicating the method and to access the data.* Once method and data are available, a statistical analysis on the behavior of deterministic algorithms is inherently less problematic than on erratic humans. Sturm (2012) showed that even if data and method are available, replication can be challenging if the problem is ill-defined and annotated data contains inconsistencies.

Even if there is little doubt on the accuracy of re-

[1] The music meta-data service Musicbrainz, for example, uses Chromaprint to assign a unique identifier to a recording, based on the audio content.

ported results, the underlying need for replication remains. First of all, it checks if the problem is well-defined. Secondly, it tests if the method is described well and in fine enough detail. Thirdly,it tests if the data used are described well and accessible. And finally results are also confirmed. It serves basically to *check if proper scientific methodology is used* and solidifies the original work.

1.1 Open Science and MIR

Open Science doesn't come as a set of prescriptive rules, but rather as a set of principles centred around the concept of 'openness', with (i) theoretical, (ii) technological/practical and (iii) ethical implications. Each scientific community needs to identify how Open Science applies to its own domain, developing 'the infrastructures, algorithms, terminologies and standards required to disseminate, visualise, retrieve and re-use data' (Leonelli, 2016, p.5). A general survey on Open Science policies in the field of MIR has never been performed, so an overview of their current application and their specific declination is not clearly defined. However, the members of this community have an implicit understanding of their own methods and their common practices to spread their materials and outputs, making it possible to lay out some fixed points.

Implementing Open Science policies in their full potential would change the face of science practice as we know it. But its achievement requires a capillar change in how we understand our day-to-day research activities and how we carry them out, and right now we are in a situation where most researchers endorse openness yet 'struggle to engage in community-oriented work because of the time and effort required to format, curate, and make resources widely available' Leonelli and Ankeny (2015). At the same time, the adoption of Open Science policies is encouraged but not mandatory and the 'variety of constraints and conditions relevant to the sharing of research materials' creates 'confusion and disagreement' among researchers (Levin et al., 2016, p.130). A recent survey of biomedical researchers in the United Kingdom (Levin et al., 2016) identified 9 external factors that affect the practice of Open Science, including the existence (or lack) of repositories and databases for data, materials, software and models; the credit system in academic research; models and guidelines for intellectual property; collaborations with industrial partners, as well as attempts at commercialization and the digital nature of research. These constraints are generally applicable across scientific domains, thus including MIR – where the aspect of commercialization emerges much earlier in the research workflow, at the level of music collections that need to be purchased.

So thinking of Open Science in MIR, where systematic support of reproducibility is but one of the possible applications, is an invitation to think about openness in relation to 'all components of research, including data, models, software, papers, and materials such as experimental samples' (Levin et al., 2016, p.132). An important and cross-domain side aim of Open Science is also to show the importance of 'encouraging critical thinking and ethical reflection among the researchers involved in data processing practices' (Leonelli, 2016, p.3). Open Science is not only about materials and platforms, but about people: the 'social' is not merely 'there' in science: 'it is capitalised upon and upgraded to become an instrument of scientific work' (Knorr-Cetina, 1999, p.29).

1.2 Replicating an acoustic fingerprinting system

This work replicates an acoustic fingerprinting system. This makes it one of the very few reported replication articles in Music Information Retrieval. Sturm and Noorzad (2012); Sturm (2012) also replicated MIR systems. They replicated two musical genre classification systems to critically review the systems and to challenge the reported high performances. Our aim is to highlight the reproducibility aspects of a milestone acoustic fingerprinting paper and to provide an illustration of good research practices. In doing so we will also provide an implementation to the research community, as well as solidifying the original acoustic fingerprinting research.

An acoustic fingerprint is a condensed representation of audio that can be matched reliably and quickly

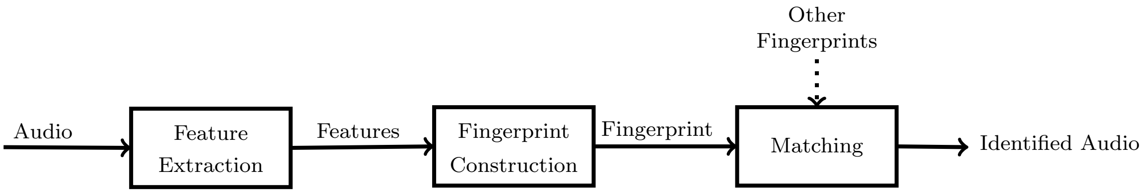

Figure 1: **A generalized audio fingerprinter scheme**. Audio is fed into the system, features are extracted and fingerprints constructed. The fingerprints are consecutively compared with a database containing the fingerprints of the reference audio. The original audio is either identified or, if no match is found, labeled as unknown.

with a large set of fingerprints extracted from reference audio. The general acoustic fingerprinting system process is depicted in Figure 1. A short query is introduced in the system. Fingerprints are extracted from the query audio and subsequently compared with a large set of fingerprints in the reference database. Finally, either a match is found or it is reported that it is not present in the database. Such acoustic fingerprint systems have many usecases such as digital rights management, identifying duplicates (Cotton and Ellis, 2010; Six et al., 2018), audio synchronization (Six and Leman, 2015) or labeling untagged audio with meta-data (Bressan et al., 2017).

The requirements for an acoustic fingerprinting system are described by Cano et al. (2005). They need to be granular, robust, reliable and economic in terms of storage requirements and computational load while resolving a query. Granular means that only a short fragment is needed for identification. Robustness is determined by various degradations a query can be subjected to while remaining recognizable. Degradations can include additional noise, low-quality encoding, compression, equalization, pitch-shifting and time-stretching. The ratios between true/false positives/negatives determine the reliability. To allow scaling to millions of reference items, an economy in terms of storage space is needed. Finally, resolving a query needs to be economic in terms of computational load. The weight of each requirement can shift depending on the context: if only about a hundred items end up in the reference database, the low storage space requirement is significantly relaxed.

Acoustic fingerprinting is a well researched MIR topic and over the years several efficient acoustic fingerprinting methods have been introduced (Herre et al., 2002; Wang, 2003; Haitsma and Kalker, 2002; Ellis et al., 2011; Allamanche, 2001; Coover and Han, 2014). These methods perform well even with degraded audio quality and with industrial sized reference databases. Some systems are able to recognize audio even when pitch-shifts are present (Fenet et al., 2011; Bellettini and Mazzini, 2008; Ramona and Peeters, 2013; Ouali et al., 2014) but without allowing for time-scale modification. Other systems are designed to handle both pitch and time-scale modification at the same time for small (Zhu et al., 2010; Malekesmaeili and Ward, 2013) datasets, or relatively large ones (Wang and Culbert, 2009; Six and Leman, 2014; Sonnleitner and Widmer, 2016).

This work replicates and critically reviews an acoustic fingerprinting system by Haitsma and Kalker (2002). The ISMIR proceedings article is from 2002 and it is elaborated upon by an article in the Journal of New Music Research (Haitsma and Kalker, 2003). The paper was chosen for several reasons:

1. It is widely cited: the ISMIR paper is cited more than 750 times and more than 250 times since 2013 according to Google Scholar. This indicates that it is *relevant* and still relevant today. A recent study, for example, improved the system by replacing the FFT with a filter bank (Plapous et al., 2017). Another study (Coover and Han, 2014) improved its robustness against noise.

2. It is a paper that has a very *prototypical* struc-

ture which presents and evaluates a MIR system. The system, in this case, is an acoustic fingerprinting system. Replicating this work, in other words, should be similar to replicating many others.

3. The described algorithm and the evaluation method are *only moderately complex and self-contained*. They only depend on regularly available tools or methods. Note that this reason is symptomatic of the reproducibility problem: some papers are borderline impossible to replicate.

1.3 Contributions

The contributions of this article are either generally applicable or specifically about the replicated work. The specific contributions are the verification of the results described by Haitsma and Kalker (2002) and a solidification of the work. A second contribution lies in a publicly available, verifiable, documented implementation of the method of that paper[2]. Another contribution is the reproducible evaluation framework. The more general contributions are a further problematization of reproducibility in MIR and guidelines to make MIR work reproducible.

The paper continues with introducing the method that is replicated and the problems encountered while replicating it. Subsequently the same is done for the evaluation. To ameliorate problems with respect to replicability in the original evaluation, an alternative evaluation is proposed. The results are compared and finally a discussion follows where guidelines are proposed.

2 Fingerprint extraction and search strategy

As with most acoustic fingerprinting systems this method consists of a fingerprint extraction step and a search strategy. In the terminology of Figure 1 this would be the feature extraction/fingerprint construction step and the matching step.

2.1 Fingerprint extraction

The fingerprint extraction algorithm is described in more detail in section 4.2 of Haitsma and Kalker (2002) but is summarized here as well. First of all the input audio is resampled to 5500Hz. On the resampled signal a Hamming windowed FFT with a length of 2048 samples is taken every 64 samples - an overlap of 96.7%. In the FFT output only 33 logarithmically spaced bins between $300Hz$ to $2000Hz$ in the magnitude spectrum are used. The energy of frequency band m at frame index n is called $E(n,m)$, it is determined by summing the energy of each FFT bin whitin each band. Finally, a fingerprint $F(n,m)$ is constructed using the $E(n,m)$ with the following formula:

$$v = \text{E(n,m)-E(n,m+1)-(E(n-1,m)-E(n-1,m+1))}$$

$$F(n,m) = \begin{cases} 1 \text{ if } v > 0 \\ 0 \text{ if } v \leq 0 \end{cases}$$

Since the last frequency band is discarded - there is no $m+1$ for the last band - only 32 of the original 33 values remain. Every FFT frame is reduced to a 32bit word. Figure 2 shows a three second audio fragment comparing the original (Figure 2-a) with a 128kb/s CBR MP3 encoded version. Figure 2-b shows the difference between the two. Figure 2-c shows the distance measure for this acoustic fingerprinting system: the number of places where the two binary words differ (in red in Figure 2). This distance measure is also known as the Hamming distance or the bit error rate BER.

Figure 3 provides a bit more insights into the BER in two cases. In the first case a high quality query, a 128 kb/s CBR encoded MP3, is compared with the reference and only a small number of bits change. Note that there are quite a few places where the BER is zero. The other case uses a low quality GSM codec. The BER, in this case, is always above zero.

Replication

The original paper includes many details about chosen parameters. It defines an FFT size, window function and sample rate, which is a good start. Unfortunately the parameters are not consistently used

[2] Available at `https://github.com/JorenSix/Panako`

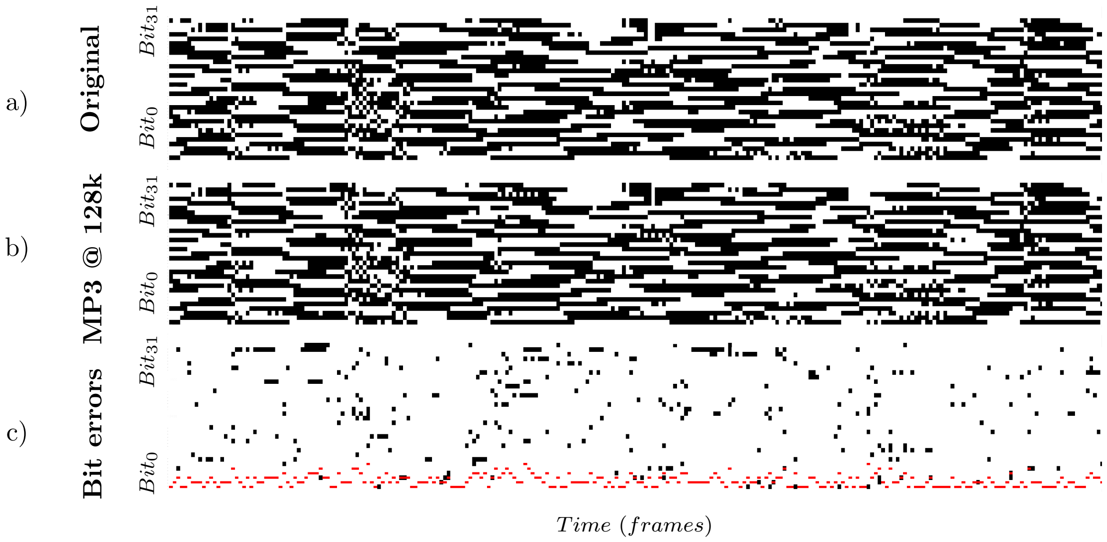

Figure 2: a) Fingerprint block of original music clip, (b) fingerprint block of a compressed version, (c) the difference between a and b showing the bit errors in black. The hamming distance or the number of bit errors is indicated in red.

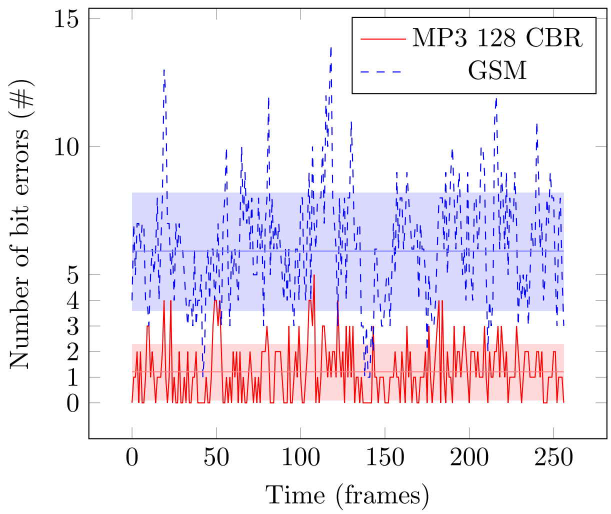

Figure 3: Bit errors per fingerprint for the 128kb/s CBR encoded MP3 and the GSM encoded version of the same three seconds of audio. Both are compared to the original uncompressed version. The average and standard deviation are indicated.

throughout the paper. Twice 11.6ms is reported as the FFT step size and twice 11.8ms. In the replicated implementation an 11.6ms step size is used (11.636ms = 64/5500Hz)[3]. While small changes to these parameters probably only have limited effect on the overall performance of the system it is exactly *these ambiguities that make an exact replication impossible*. The replication, in other words, includes various assumptions on the original system which may be false. However, even under these conditions the general behavior of the system may still remain intact. Results are expected to differ but only up until a certain unknown degree.

Furthermore, consistent textual description of an algorithm always leaves some wiggle room for different interpretations (Peng, 2011). Only if source code is available with details on which system - software and hardware - the evaluation is done can an exact replication become feasible. The source code could

[3] Note that a sample rate of 5.5kHz is used and not, as reported in the original paper, 5kHz, to end up at the correct step size. The follow up journal article Haitsma and Kalker (2003) does mention 5.5Hz

also include bugs that perhaps have an effect on the results. Bugs will, by definition, not be described as such in a textual description.

This strengthens the case that source code should be an integral part of a scientific work. If interested in further details of the new implementation readers are referred to the source code in the supplementary material[4]. There, it becomes clear at which precision the FFT was calculated, how exactly downsampling was done, the precision of the windowing function, and so forth.

2.2 Search strategy

The basic principle of the search strategy is a nearest neighbor search in Hamming space. For each fingerprint extracted from a query, a list of near neighbors is fetched which ideally includes a fingerprint from the matching audio fragment. The actual matching fragment will be present in most lists of near neighbors. The details of the search strategy are much less critical then the parameters of the fingerprint extraction step. As long as a nearest neighbor search algorithm is implemented correctly the only difference will be the speed at which a query is resolved.

The search strategies' main parameter is the supported Hamming distance. With an increased Hamming distance d more degraded audio can be retrieved but the search space quickly explodes. For b the number of bits the search space equals:

$$\sum_{k=d}^{k-2|k\geq 0} \frac{b!}{k!(b-k)!} = \sum_{k=d}^{k-2|k\geq 0} \binom{b}{k} \qquad (1)$$

The search strategy from the original work keeps track of which bits of a fingerprint are uncertain. The uncertain bits are close to the threshold. It assigns each bit with a value from 1 to 32 that describes confidence in the bit, with 1 being the least reliable bit and 32 the most reliable. Subsequently, a search is done for the fingerprint itself and for fingerprints which are permutations of the original with one or more uncertain bits toggled. To strike a balance between performance and retrieval rate the number of bits need to be chosen. If the three least reliable bits are toggled, this generates 2^3 permutations. This is much less than flipping 3 bits randomly in the 32bit fingerprint:

$$\sum_{k=3}^{k-2|k\geq 0} \binom{33}{3} = 5489$$

Once a matching fingerprint is found the next step is to compare a set of fingerprints of the query with the corresponding set of fingerprints in the reference audio. The Hamming distance for each fingerprint pair is calculated. If the sum of the distances is below a threshold then it is declared a match, otherwise the search continues until either a match is found or until the query is labeled as unknown. The parameters are determined experimentally in the original work: 256 fingerprints are checked and the threshold for the Hamming distance is 2867bits. So from a total of $256 \times 32bits$, 2867 or about 35% are allowed to be different.

Replication

The implementation is done with two hash tables. A lookup table with fingerprints as key and a list of $(identifier, offset)$ pairs as value. The identifier refers uniquely to a track in the reference database. The offset points precisely to the time at which the fingerprint appears in that track. The second hash table has an identifier as key and an array of fingerprints as value. Using the offset, the index in the fingerprint array can be determined. Subsequently, the previous 256 fingerprints from the query can be compared with the corresponding fingerprints in the reference set and a match can be verified.

Implementing this search strategy is relatively straightforward.

3 Evaluation

The evaluation of the system is done in two ways. First we aim to replicate the original evaluation and match the original results as closely as possible to validate the new implementation and the original work. The original evaluation is not easily replicated since

[4] Also available at `https://github.com/JorenSix/Panako`

it uses copyrighted evaluation audio with ambiguous descriptions, a data set that is not available or described and modifications that are only detailed up until a certain degree.

The second evaluation is fully replicable: it uses freely available evaluation audio, a data set with creative commons music and modifications that are encoded in scripts. Interested readers are encouraged to replicate the results in full.

3.1 Replication of the original evaluation

The evaluation of the original system is done on four short excerpts from commercially available tracks: ‘O Fortuna’ by Carl Orff, ‘Success has made a failure of our home’ by Sinead O’Connor, ‘Say what you want’ by Texas and ‘A whole lot of Rosie’ by AC/DC. Unfortunately it fails to mention how long these excerpts were or where in the song they originate. The selection does have an effect on performance. If a part with little acoustic information is selected versus a dense part different results can be expected. It also fails to mention which edition, version or release is employed which is problematic with the classical piece for which many varying performances exist. The paper also mentions a reference database of 10 000 tracks but fails to specify which tracks it contains. The fact that only one excerpt from each song is used for evaluation makes the selection critical which is problematic by itself. Reporting an average performance with standard deviations would have been more informative.

To evaluate the robustness of the system each short excerpt is modified in various ways. The modifications to the query are described well but there is room for improvement. For example, it is not mentioned how time-scale modification is done: there are different audible artifacts - i.e. different results - when a time or frequency domain method for time-scale modification is used. The description of the echo modification seems to have been forgotten while dry, wetness or delay length parameters are expected to have a large effect on results.

To summarize: essential information is missing to replicate the results exactly. The next best thing is to follow the basic evaluation method which can be replicated by following various clues and assumptions. To this end the previously mentioned four tracks were bought from a digital music store (7digital, see table 1). Two were available in a lossless format and two in a high quality MP3 format (320 kb/s CBR). The test data set can not be freely shared since commercial music is used which, again, hinders replicability.

The original evaluation produces two tables. The first documents the bit error rates (BER, Table 2). It compares the fingerprints extracted from a reference recording with those of modified versions. If all bits are equal the error rate is zero. If all bits are different then the error rate is one. Comparison of random fingerprints will result in a bit error rate of around 0.5. The original article suggests that 256 fingerprints (about three seconds) are compared and the average is reported. Experimentally the original article determines that a BER of 0.35 or less is sufficient to claim that two excerpts are the same with only a very small chance of yielding a false positive. The BER evaluation has been replicated but due to the fact that the excerpts are not identical and the modifications also deviate slightly, the replicated BER values differ. However, if the original and replicated results are compared using a Pearson correlation there is a very strong linear relation $r(58) = 0.92, p < 0.001$. This analysis suggests that the system behaves similarly for the various modifications. The analysis left out the white noise condition, which is an outlier. The replicated modification probably mixed more noise into the signal than the original. Some modifications could not be successfully replicated either because they are not technically relevant (cassette tape, real media encoding) or the method to do the modification was unclear (GSM C/I).

A second table (Table 3) shows how many of 256 fingerprints could be retrieved in two cases. The first case tries to find only the exact matches in the database. The reported number shows how many of the 256 fingerprints point to the matching fingerprint block in the database. If all fingerprints match, the maximum (256) is reported. In the second case the 10 most unreliable bits are flipped resulting in 1024 fingerprints which are then matched with the database.

Identifier	ISRC	AcoustID	Track	Format
56984036		3af00f3a-afc8-4b62-8eff-dacb7d7245c9	Sinead	320kbs MP3
52740482		b03406c9-1b14-427e-b4fa-16029b8a72cc	ACDC	16-bit/44.1kHz FLAC
122965	GBF089607481	92f4e392-a36e-47c8-bfee-b553b0c0e0ad	Texas	320kbs MP3
5917942	DEF056730100	99eb4952-9a72-4811-9b1e-f8c8ab737e9f	Orff	16-bit/44.1kHz FLAC

Table 1: Tracks bought from 7digital music store with 7digital identifier and format information. The ISRC (International Standard Recording Code) and AcoustID [6] fingerprint are provided as well.

	Texas		Sinead		Orff		AC/DC	
Modification	**Original**	**Replication**	**Original**	**Replication**	**Original**	**Replication**	**Original**	**Replication**
MP3@128Kbps	0.081	0.055	0.085	0.077	0.078	0.056	0.084	0.035
MP3@32Kbps	0.096	0.097	0.106	0.115	0.174	0.100	0.133	0.089
Real@20Kbps	0.159	/	0.138	/	0.161	/	0.21	/
GSM	0.168	0.194	0.144	0.211	0.16	0.217	0.181	0.187
GSM C/I = 4dB	0.316	/	0.247	/	0.286	/	0.324	/
All-pass filtering	0.018	0.020	0.015	0.032	0.019	0.033	0.027	0.010
Amp. Compr.	0.113	0.010	0.07	0.027	0.052	0.033	0.073	0.014
Equalization	0.066	0.025	0.045	0.024	0.048	0.023	0.062	0.013
Echo Addition	0.139	0.132	0.148	0.145	0.157	0.118	0.145	0.109
Band Pass Filter	0.024	0.031	0.025	0.034	0.028	0.030	0.038	0.017
Time Scale +4%	0.2	0.279	0.183	0.283	0.202	0.302	0.206	0.301
Time Scale 4%	0.19	0.263	0.174	0.277	0.207	0.281	0.203	0.294
Linear Speed +1%	0.132	0.189	0.102	0.193	0.172	0.214	0.238	0.181
Linear Speed -1%	0.26	0.177	0.142	0.199	0.243	0.201	0.196	0.177
Linear Speed +4%	0.355	0.434	0.467	0.461	0.438	0.551	0.472	0.470
Linear Speed -4%	0.47	0.425	0.438	0.500	0.464	0.510	0.431	0.464
Noise Addition	0.011	0.042	0.011	0.122	0.009	0.273	0.036	0.027
Resampling	0	0.000	0	0.000	0	0.004	0	0.000
D/A A/D	0.111	/	0.061	/	0.088	/	0.076	/

Table 2: Replication of bit error rates (BER) for different kinds of signal degradations. The original results and replicated results are reported.

In both cases only one correct hit is needed to identify an audio excerpt.

Again, the original results are compared with the replicated results with a Pearson correlation. The exact matching case shows a strong linear correlation $r(62) = 0.66, p < 0.001$ and the case of 10 flipped bits show similar results $r(62) = 0.67, p < 0.001$. This suggests that the system behaves similarly considering that the audio excerpts, the modifications and implementation include differences and various assumptions had to be made.

3.2 A replicable evaluation

The original evaluation has several problems with respect to replicability. It uses commercial music but fails to mention which exact audio is used both in the reference database as for the evaluation. The process to generate modification is documented but still leaves room for interpretation. There are also other problems: The evaluation depends on the selection of only four audio excerpts.

The ideal acoustic fingerprinting system evaluation depends on the use-case. For example, the evaluation method described by Ramona et al. (2012) focuses mainly on broadcast monitoring and specific modifications that appear when broadcasting music over the radio. The SyncOccur corpus (Ramona and Peeters, 2013) also focuses on this use-case. An evaluation of an acoustic fingerprinting system for DJ-set monitoring (Sonnleitner et al., 2016), low-power consumption on smartphones (y Arcas et al., 2017) or sample identification (Van Balen et al., 2012) needs another approach. These differences in focus lead to a wide variety of evaluation techniques for systems which makes them hard to compare. The *replicable* evaluation described here evaluates a fingerprint system for (re-encoded) duplicate detection with simple degradations[7]. Modern fingerprinting systems are more robust to noise and pitch/time scaling. Here, however, we *focus on what makes an evaluation replicable.* The evaluation presented below requires only open source software and is similar to the procedure used already by Sonnleitner and Widmer (2016) and Six and Leman (2014).

The evaluation is done as follows. Using a script, available as supplementary material, 10,100 creative commons licensed musical tracks are downloaded from Jamendo[8]. 10,000 of these tracks are added to the reference database. The remaining 100 are not. The script provides a list of Jamendo track identifiers that are used in the reference database. Using another script, 1100 queries are selected at random[9]. The queries are divided into 1000 from the items that are in the database and 100 from items that are not present in the reference database. This is to check true negatives. A query is three seconds long and starts at 30 seconds into the song. Each query is modified automatically using the modifications described above. This modification process is also automatized with SoX, a command line audio editor. Subsequently, these queries are matched with the reference database. The main parameters of the evaluation are the amount of unreliable bits to flip and the threshold when a match is accepted. The number of unreliable bits to flip was set to 10. If less than 2867 bits are different between 256 subsequent 32bits fingerprints then the match is accepted.

Once the results are available each query is checked if it is either a true positive TP, false positive FP, true negative TN or false negative FN. Next to TP, FP, TN and FN sensitivity, specificity, precision and accuracy are calculated as well.

Table 4 summarizes the results. As expected the system's specificity and precision is very high. The few cases where a false positive is reported is due to audio duplicates in the reference database. The reference database does contain a few duplicate items where audio is either completely the same or where parts of another track are sampled. Note that the evaluation is done on the track level, the time offset is not taken into account. Since exact repetition is not uncommon, especially in electronic music, a query can be found at multiple, equally correct time offsets.

[7] If complex but repeatable audio degradations are needed the Audio Degradation MatLab toolbox by Mauch and Ewert (2013) can be of interest.

[8] Jamendo is a music sharing service with music shared under specific non-exclusive licenses such as Creative Commons.

[9] The random function uses a fixed seed so that the evaluation can be either repeated exactly or, when given a different seed, verified with another set

	Orff		**Sinead**		**Texas**		**AC/DC**	
Modification	**Original**	**Replication**	**Original**	**Replication**	**Original**	**Replication**	**Original**	**Replication**
MP3@128Kbps	17, 170	150, 226	20, 196	59, 111	23, 182	94, 166	19, 144	144, 207
MP3@32Kbps	0, 34	44, 123	10, 153	14, 63	13, 148	20, 56	5, 61	29, 87
Real@20Kbps	2, 7	/	7, 110	/	2, 67	/	1, 41	/
GSM	1, 57	2, 6	2, 95	0, 1	1, 60	0, 5	0, 31	4, 16
GSM C/I = 4dB	0, 3	/	0, 12	/	0, 1	/	0, 3	/
All-pass filtering	157, 240	170, 244	158, 256	161, 226	146, 256	166, 251	106, 219	191, 245
Amp. Compr.	55, 191	145, 222	59, 183	98, 156	16, 73	169, 247	44, 146	183, 241
Equalization	55, 203	161, 236	71, 227	220, 126	34, 172	126, 193	42, 148	171, 227
Echo Addition	2, 36	53, 70	12, 69	37, 73	15, 69	68, 112	4, 52	73, 102
Band Pass Filter	123, 225	169, 237	118, 253	149, 193	117, 255	110, 186	80, 214	159, 241
Time Scale +4%	6, 55	43, 72	7, 68	53, 54	16, 70	57, 123	6, 36	66, 118
Time Scale 4%	17, 60	57, 107	22, 77	53, 57	23, 62	54, 118	16, 44	60, 108
Linear Speed +1%	3, 29	2, 6	18, 170	2, 16	3, 82	3, 22	1, 16	8, 35
Linear Speed -1%	0, 7	0, 8	5, 88	2, 16	0, 7	1, 22	0, 8	4, 16
Linear Speed +4%	0, 0	0, 0	0, 0	0, 0	0, 0	0, 0	0, 1	0, 0
Linear Speed -4%	0, 0	0, 0	0, 0	0, 0	0, 0	0, 0	0, 0	0, 0
Noise Addition	190, 256	30, 73	178, 255	0, 9	179, 256	23, 101	114, 255	99, 167
Resampling	255, 256	253, 256	255, 256	239, 256	254, 256	254, 256	254, 256	253, 256
D/A A/D	15, 149	/	38, 229	/	13, 114	/	31, 145	/

Table 3: Replication of hits in the database for different kinds of signal degradations. First number indicates the hits for using only the 256 sub-fingerprints to generate candidate positions. Second number indicates hits when 1024 most probable candidates for every sub-fingerprint are also used

	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy	Precision	Avg. dist. (bits)
MP3@128Kbps	90.53%	9.18%	0.09%	0.19%	99.79%	98.99%	99.72%	99.90%	4.72 ± 1.64
MP3@32Kbps	89.97%	9.18%	0.19%	0.66%	99.28%	98.00%	99.16%	99.79%	5.73 ± 1.72
All-pass filtering	90.35%	9.18%	0.00%	0.47%	99.48%	100.00%	99.53%	100.00%	5.09 ± 1.72
Amp. Compr.	90.44%	9.18%	0.00%	0.37%	99.59%	100.00%	99.63%	100.00%	5.26 ± 1.79
Band Pass Filter	90.63%	9.18%	0.09%	0.09%	99.90%	98.99%	99.81%	99.90%	5.13 ± 1.75
Echo Addition	86.14%	9.27%	0.19%	4.40%	95.14%	98.02%	95.41%	99.78%	7.12 ± 1.53
Equalization	90.63%	9.18%	0.00%	0.19%	99.79%	100.00%	99.81%	100.00%	5.25 ± 1.78
GSM	42.92%	9.28%	0.19%	47.61%	47.41%	98.02%	52.20%	99.57%	9.02 ± 1.17
Resampling	90.43%	9.19%	0.09%	0.28%	99.69%	98.99%	99.62%	99.90%	4.95 ± 1.68
Linear Speed -4%	0.00%	9.27%	0.00%	90.73%	0.00%	100.00%	9.27%	/	/
Linear Speed -1%	75.66%	9.27%	0.19%	14.89%	83.56%	98.02%	84.93%	99.75%	8.41 ± 1.46
Linear Speed +1%	79.40%	9.27%	0.28%	11.05%	87.78%	97.06%	88.67%	99.65%	7.65 ± 1.41
Linear Speed +4%	0.00%	9.27%	0.00%	90.73%	0.00%	100.00%	9.27%	/	/
Time Scale -4%	76.50%	9.27%	0.28%	13.95%	84.58%	97.06%	85.77%	99.63%	10.20 ± 0.88
Time Scale +4%	88.30%	9.27%	0.19%	2.25%	97.52%	98.02%	97.57%	99.79%	9.72 ± 1.00
Noise Addition	87.83%	9.27%	0.19%	2.72%	97.00%	98.02%	97.10%	99.79%	5.60 ± 1.98

Table 4: Results on a dataset of 10k songs with 1000 queries per modification. The average Hamming distance between a modified fingerprint of 32 bits and the matching reference is reported ± one standard deviation.

If the system returns the correct track identifier with an unexpected offset then it is still counted as a true positive.

The sensitivity and accuracy of the system goes down when the average bit errors per fingerprint approaches the threshold of 10 erroneous bits. True positives for GSM encoded material are only found about half of the time. The average Hamming distance in bits for queries with a changed time scale of ±4% is higher than the GSM encoded queries while accuracy is much higher. This means that for the GSM encoded material the reliability information is not reliable: the 10 least reliable bits are flipped but still the original fingerprint is not found for about half of the queries.

There are some discrepancies between these results and the reported results in the original study. The average Hamming distance between queries and reference is higher in the new evaluation. This is potentially due to the use of 128kbs MP3's during the evaluation. The original material is decoded to store in the reference database and the queries are re-encoded after modification. Another discrepancy is related to the GSM encoded queries: the original results seem to suggest that all GSM encoded queries would yield a true positive (see table 3). This was not achieved in the replication. Whether this is due to incorrect assumptions, different source material, the evaluation method or other causes is not clear.

4 Discussion

As statistical comparison showed, the replicated system behaves generally in a similar way as the originally described system. On top of that an alternative, reproducible, evaluation showed that following the system's design allows for functional acoustic fingerprinting. There are however *unexplained discrepancies* between both systems especially concerning the GSM modification. It is worrisome that it is impossible to pinpoint the source of these discrepancies since neither the original evaluation material, evaluation method, nor implementation are available. While there is no guarantee that the replication is bug free, at least the source can be checked.

All in all, the results are quite similar to the original. As stated in the introduction replication of results should be expected to pose no problem. It is, however, the replication of methods and accessibility of data that makes replication prohibitively *time-consuming*. This could be alleviated with releasing research code and data. While the focus of the MIR community should remain on producing novel techniques to deal with musical information and not on producing end-user ready software, it would be beneficial for the field to keep sustainable software aspects in mind when releasing research prototypes. Aspects such as those identified by Jackson et al. (2011) where a distinction is made between usability (documentation, installability,...) and maintainability (identity, copyright, accessibility, interoperability,...).

5 Conclusion

Intellectual property rights, copyrights on music and a lack of incentive pose a problem for reproducibility of MIR research work. There are, however, ways to deal with these limiting factors and foster reproducible research. We see a couple of work-arounds and possibilities, which are described below.

As universities are striving more and more for open-access publications there could be a similar movement for data and code. After all, it makes little sense to publish only part of the research in the open (the textual description) while keeping code and data behind closed doors. Especially if the research is funded by public funds. In Europe, there is an ambition to make all scientific articles freely available by 2020 and to achieve optimal reuse of scientific data[10], though research software seems to have been forgotten in this directive. A change in attitude towards releasing more software for research institutions and public funded universities is needed. A good starting point would be updating publication policies to include software together with a clear stance on **intellectual property rights**.

[10] All European scientific articles to be freely accessible by 2020 - Europe makes a definitive choice for open access by 2020 - 27 May 2016 - Michiel Hendrikx

Copyrights on music make it hard to share music freely. We see two ways to deal with this:

1. *Pragmatic vs Ecological* or Jamendo *vs* iTunes. There is a great deal of freely available music published under various creative commons licenses. Jamendo for example contains half a million cc-licensed tracks which are uniquely identifiable and can be download via an API. Much of the music that can be found there is recorded at home with limited means. This means that systems can behave slightly differently on the Jamendo set when compared with a set of commercial music. What is gained in pragmatism is perhaps lost in ecological validity. Whether this is a problem depends very much on the research question at hand. In the evaluation proposed here Jamendo was used (similarly to Sonnleitner and Widmer (2016); Six and Leman (2014)) since it does offer a large variability in genres and is representative for this use-case.

2. *Audio vs Features.* Research on features extracted from audio does not need audio itself: if the features are available this can suffice. There are two large sets of audio features. The million song data set (Bertin-Mahieux et al., 2011) and Acousticbrainz (Porter et al., 2015). Both ran feature extractors on millions of commercial tracks and have an API to query or download the data. Unfortunately the source of the feature extractors used in the Million Song data set are not available and only described up until a certain level of detail which makes it a black box and, in our eyes, unfit for decent reproducible science. Indeed, due to internal reorganizations and mergers the API and the data have become less and less available. The science built on the million song dataset is on shaky ground. Fortunately, Acousticbrainz is completely transparent. It uses well documented, open source software (Bogdanov et al., 2013) and the feature extractors are reproducible. The main shortcoming of this approach is that only a curated set of features is available. If another feature is needed, then you are out of luck. Adding a feature is far from trivial, since even Acousticbraiz has no access to all audio: they rely on crowdsourced feature extraction.

Providing an **incentive** for researchers to make their research reproducible is hard. This requires a mentality shift. Policies by journals, conference organizers and research institutions should gradually change to require reproducibility. There are a few initiatives to foster reproducible research, specifically for music informatics research. The 53rd Audio Engineering Society (AES) conference had a price for reproducibility. ISMIR 2012 had a tutorial on *'Reusable software and reproducibility in music informatics research'* but structural attention for this issue at ISMIR seems to lack. There is, however, a yearly workshop organized by Queen Mary University London (QMUL) on 'Software and Data for Audio and Music Research'. It 'include talks on issues such as robust software development for audio and music research, **reproducible research in general**, management of research data, and open access'.[11]. At QMUL there seems to be continuous attention to the issue and researchers are trained in software craftsmanship[12]

In this article we problematized reproducibility in MIR and illustrated this by replicating an acoustic fingerprinting system. While similar results were obtained there are unexplained and *unexplainable discrepancies* due to the fact that the original data, method and evaluation is only partly available and assumptions need to be made. We proposed an alternative, reproducible, evaluation and extrapolated general guidelines aiming to improve reproducibility of MIR research in general.

Acknowledgements

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant

[11] `http://soundsoftware.ac.uk/soundsoftware2014` - March 2017

[12] They also host a repository for software dealing with sound at `http://soundsoftware.ac.uk`.

agreement No. 703937 and partially supported by an FWO Methusalem project titled *Expressive Music Interaction.* The authors would also like to thank dr. William Harwood for proofreading this article.

References

Allamanche, E. (2001). Content-based identification of audio material using MPEG-7 low level description. In *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR 2001).*

Bellettini, C. and Mazzini, G. (2008). Reliable automatic recognition for pitch-shifted audio. In *Proceedings of 17th International Conference on Computer Communications and Networks (ICCCN 2008)*, pages 838–843. IEEE.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011).*

Bogdanov, D., Wack, N., Gmez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *Proceedings of the 14th International Symposium on Music Information Retrieval (ISMIR 2013)*, pages 493–498.

Bressan, F., Six, J., and Leman, M. (2017). Applications of duplicate detection: linking meta-data and merging music archives. The experience of the IPEM historical archive of electronic music. In *Proceedings of 4th International Digital Libraries for Musicology workshop (DLfM 2017)*, page submitted, Shanghai (China). ACM Press.

Cano, P., Batlle, E., Kalker, T., and Haitsma, J. (2005). A review of audio fingerprinting. *The Journal of VLSI Signal Processing*, 41:271–284.

Collaboration, O. S. et al. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Coover, B. and Han, J. (2014). A power mask based audio fingerprint. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1394–1398.

Cotton, C. V. and Ellis, D. P. W. (2010). Audio fingerprinting to identify multiple videos of an event. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2386–2389. IEEE.

Ellis, D., Whitman, B., and Porter, A. (2011). Echoprint - an open music identification service. In *Proceedings of the 12th International Symposium on Music Information Retrieval (ISMIR 2011).*

Fenet, S., Richard, G., and Grenier, Y. (2011). A Scalable Audio Fingerprint Method with Robustness to Pitch-Shifting. In *Proceedings of the 12th International Symposium on Music Information Retrieval (ISMIR 2011)*, pages 121–126.

Fischinger, T. (2013). Preface by the guest editor of the special issue of musicae scientiae on replication in music psychology.

Haitsma, J. and Kalker, T. (2002). A highly robust audio fingerprinting system. In *ISMIR 2002, 3rd International Conference on Music Information Retrieval, Paris, France, October 13-17, 2002, Proceedings.*

Haitsma, J. and Kalker, T. (2003). A highly robust audio fingerprinting system with an efficient search strategy. *Journal of New Music Research*, 32(2):211–221.

Herre, J., Hellmuth, O., and Cremer, M. (2002). Scalable robust audio fingerprinting using MPEG-7 content description. In *Multimedia Signal Processing, 2002 IEEE Workshop on*, pages 165–168. IEEE.

Jackson, M., Crouch, S., and Baxter, R. (2011). Software evaluation: criteria-based assessment. *Software Sustainability Institute.*

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge.* Harvard University Press.

Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A*, 374.

Leonelli, S. and Ankeny, R. A. (2015). Repertoires: How to transform a project into a research community. *BioScience*, 65(7):701–708.

Levin, N., Leonelli, S., Weckowska, D., Castle, D., and Dupré, J. (2016). How do scientists define openness? exploring the relationship between open science policies and research practice. *Bulletin of Science, Technology and Society*, 36(2):128–141.

Malekesmaeili, M. and Ward, R. K. (2013). A local fingerprinting approach for audio copy detection. *Computing Research Repository (CoRR)*, abs/1304.0793.

Mauch, M. and Ewert, S. (2013). The audio degradation toolbox and its application to robustness evaluation. In *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013 2013*, pages 83–88.

Mesirov, J. P. (2010). Accessible reproducible research. *Science*, 327(5964):415–416.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., and Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242):1422–1425.

Ouali, C., Dumouchel, P., and Gupta, V. (2014). A robust audio fingerprinting method for content-based copy detection. In *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on*, pages 1–6. IEEE.

Pashler, H. and Wagenmakers, E.-J. (2012). Editors introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530.

Peeters, G. and Fort, K. (2012). Towards a (Better) Definition of the Description of Annotated MIR Corpora. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012*, pages 25–30. Citeseer.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226–1227.

Plapous, C., Berrani, S.-A., Besset, B., and Rault, J.-B. (2017). A low-complexity audio fingerprinting technique for embedded applications. *Multimedia Tools and Applications*, pages 1–20.

Porter, A., Bogdanov, D., Kaye, R., Tsukanov, R., and Serra, X. (2015). Acousticbrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval Conference (ISMIR15)*.

Ramona, M., Fenet, S., Blouet, R., Bredin, H., Fillon, T., and Peeters, G. (2012). A public audio identification evaluation framework for broadcast monitoring. *Applied Artificial Intelligence*, 26(1-2):119–136.

Ramona, M. and Peeters, G. (2013). AudioPrint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In *Proceedings of the 2013 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2013)*, pages 818–822.

Six, J., Bressan, F., and Leman, M. (preprint – 2018). Applications of duplicate detection in music archives: From metadata comparison to storage optimisation - the case of the belgian royal museum for central africa. In *Proceedings of the 13th Italian Research Conference on Digital Libraries (IRCDL 2018)*.

Six, J. and Leman, M. (2014). Panako - A Scalable Acoustic Fingerprinting System Handling Time-Scale and Pitch Modification. In *Proceedings of the 15th ISMIR Conference (ISMIR 2014)*.

Six, J. and Leman, M. (2015). Synchronizing Multimodal Recordings Using Audio-To-Audio Alignment. *Journal of Multimodal User Interfaces*, 9(3):223–229.

Sonnleitner, R., Arzt, A., and Widmer, G. (2016). Landmark-Based Audio Fingerprinting for DJ Mix Monitoring. In *ISMIR*, pages 185–191.

Sonnleitner, R. and Widmer, G. (2016). Robust quad-based audio fingerprinting. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, PP(99):1–1.

Sturm, B. L. (2012). Two systems for automatic music genre recognition: What are they really recognizing? In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 69–74. ACM.

Sturm, B. L. and Noorzad, P. (2012). On automatic music genre recognition by sparse representation classification using auditory temporal modulations. *Computer music modeling and retrieval*, pages 379–394.

Van Balen, J., Serrà, J., and Haro, M. (2012). Sample identification in hip hop music. In *International Symposium on Computer Music Modeling and Retrieval*, pages 301–312. Springer.

Wang, A. and Culbert, D. (2009). Robust and invariant audio pattern matching. US Patent 7,627,477.

Wang, A. L.-C. (2003). An industrial-strength audio search algorithm. In *Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR 2003)*, pages 7–13.

y Arcas, B. A., Gfeller, B., Guo, R., Kilgour, K., Kumar, S., Lyon, J., Odell, J., Ritter, M., Roblek, D., Sharifi, M., and Velimirovic, M. (2017). Now playing: Continuous low-power music recognition. *CoRR*, abs/1711.10958.

Zhu, B., Li, W., Wang, Z., and Xue, X. (2010). A novel audio fingerprinting method robust to time scale modification and pitch shifting. In *Proceedings of the international conference on Multimedia (MM 2010)*, pages 987–990. ACM.