Applications of duplicate detection: linking meta-data and merging music archives

The experience of the IPEM historical archive of electronic music

Federica Bressan, Joren Six, Marc Leman
IPEM, University Ghent
federica.bressan@ugent.be

This is the author version of a paper published by ACM. For a version of record, consult doi 10.1145/3144749.3144759

Abstract

This work focuses on applications of duplicate detection for managing digital music archives. It aims to make this mature music information retrieval (MIR) technology better known to archivists and provide clear suggestions on how this technology can be used in practice. More specifically applications are discussed to complement meta-data, to link or merge digital music archives, to improve listening experiences and to re-use segmentation data. The IPEM archive, a digitized music archive containing early electronic music, provides a case study.

1 Introduction

Music Information Retrieval (MIR) technologies have a lot of untapped potential in the management of digital music archives. There seem to be several reasons for this. One is that MIR technologies are simply not well known to archivists. Another reason is that it is often unclear how MIR technology can be applied in a digital music archive setting. A third reason is that considerable effort is often needed to transform a potentially promising MIR research prototype into a working solution for archivists as end-users.

In this article we focus on duplicate detection. It is an MIR technology that has matured over the last two decades for which there is usable software available. The aim of the is article is to make this technology better known to the community of archivists and to describe several applications for duplicate detection. Some of these applications might not be immediately obvious since duplicate detection is used indirectly to complement meta-data, link or merge archives, improve listening experiences and it has opportunities for segmentation. These applications are grounded in experience with working with the IPEM archive, an archive of early electronic music on tape that has been digitised twice in the past fifteen years.

2 Duplicate detection

The problem of duplicate detection is defined as follows:

> *How to design a system that is able to compare every audio fragment in a set with all other audio in the set to determine if the fragment is either unique or appears multiple times in the complete set. The comparison should be robust against various Artefacts.*

The artefacts in the definition above include noise of various sources. This includes imperfections introduced during the analog-to-digital (A/D) conversion. Artefacts resulting from mechanical defects, such as clicks from gramophone discs or magnetic tape hum. Detecting duplicates should be possible when changes in volume, compression or dynamics are introduced as well.

Over time it is almost inevitable that duplicates of the same recording end up in a digitised archive. For example, an original field recording is published on an LP, and both the LP as the original version get digitised and stored in the same lot. It is also not uncommon that an archive contains multiple copies of the same recording because the same live event was captured from two different angles (normally on the side of the parterre and from the orchestra pit), or because before the advent of digital technology, copies of degrading tapes were already being made on other tapes. Last but not least, the chance of duplicates grows exponentially when different archives or audio collections get connected or virtually merged, which is a desirable operation and one of the advantages introduced by the digital technology (see 2.4).

To summarise, using the terminology from Cano et al. (2005), the duplicate detector needs to have these requirements:

- It needs to be capable to mark duplicates without generating false positives or missing true positives. In other words **precision and recall** need to be acceptable.

- It should be capable to operate on large archives. It should be **efficient**. Efficient here means quick when resolving a query and efficient on storage and memory use when building an index.

- Duplicates should be marked as such even if there is noise or the speed is not kept constant. It should be **robust** against various modifications.

- Lookup for short audio fragments should be possible, the algorithm should be **granular**. A resolution of 20 seconds or less is beneficial.

Once such system is available, several applications are possible.

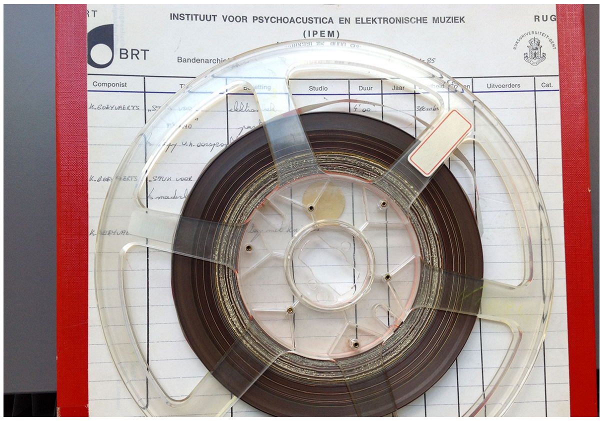

Figure 1: Open-reel tape from the IPEM archive. Content description on the box, according to the pre-defined scheme: composer, title, instrumentation, studio, duration, year, etc.

2.1 Duplicate detection for complementing meta-data

Being aware of duplicates is useful to **check or complement meta-data**. If an item has richer meta-data than a duplicate, the meta-data of the duplicate can be integrated. With a duplicate detection technology conflicting meta-data between an original and a duplicate can be resolved or at least flagged. The problem of conflicting meta-data is especially prevalent in archives with ethnic music where often there are many different spellings of names, places and titles. Naming instruments systematically can also be very challenging.

2.2 Duplicate detection to improve the listening experience

When multiple recordings in sequence are marked as exact duplicates, meaning they contain the exact same digital information, this **indicates inefficient storage use**. If they do not contain exactly the same information it is possible that either the same analog carrier was accidentally digitised twice or there are effectively two analogue copies with the same content. To **improve the listening experience** the most qualitative digitised version can be returned if requested, or alternatively to assist philological research all the different versions (variants, witnesses of the archetype) can be returned.

2.3 Duplicate detection for segmentation

It potentially solves **segmentation** issues. When an LP is digitised as one long recording and the same material has already been segmented in an other digitisation effort, the segmentation boundaries can be reused. Also duplicate detection allows to identify when different segmentation boundaries are used. Perhaps an item was not segmented in one digitisation effort while a partial duplicate is split and has an extra meta-data item – e.g. an extra title. Duplicated detection allows re-use of segmentation boundaries or, at the bare minimum, indicate segmentation discrepancies.

2.4 Duplicate detection for merging archives

Technology makes it possible to **merge or link digital archives** from different sources – e.g. the creation of a single point of access to documentation from different institutions concerning a special subject; the implementation of the "virtual re-unification" of collections and holdings from a single original location or creator now widely scattered (IFLA - Audiovisual and Multimedia Section, 2002, p.11). More and more digital music archives 'islands' are bridged by efforts such as Europeana Sounds . Europeana Sounds is a European effort to standardise metadata and link digital music archives. Table 1 shows a few of these archives. The EuropeanaConnect/DISMARC Audio Aggregation Platform provides this link and could definitely benefit from duplicate detection technology and provide a view on unique material.

If duplicates are found in one of these merged archives, all previous duplicate detection applications come into play as well. How similar is the meta-data between original and duplicate? How large is the difference in audio quality? Are both original and duplicate segmented similarly or is there a discrepancy?

2.5 Robustnsess to speed change

Duplicate detection robust to speed changes has an important added value. When playback (or recording) speed changes from analogue carriers, both tempo and pitch change accordingly. Most people are familiar with the effect of playing a 33 rpm LP at 45 rpm. But the problem with historic archives and analogue carriers is more subtle: the speed at which the tape gets digitised might not match the original recording speed, impacting the resulting pitch. Often times the exact recording speed is not indicated on the original box, but even when that is the case, it is impossible to predict with reasonable precision when the recording device was defective, inadequately operated, or when the portable recorder was slowly running out of battery. So not only it is nearly impossible to make a good estimation of the original non-standard recording speed, but it might not be a constant speed at all, it could actually fluctuate 'around' a standard speed. This is especially problematic with wax cylinders, where there are numerous speed indications but they are not systematically used – if indications are present at all.

Therefore, in the light of what has been said so far, the problem of speed fluctuation is structural and endemic in historical analogue sound archives, and cannot be easily dismissed. Hence the crucial importance of algorithms that treat this type of material to consider this problem and operate accordingly.

CNRS-CREM	21,462
Comhaltas Traditional Music Archive	14,419
Internet Archive	10,487
LMTA (DIZI)	10,234
CNRS-LARHRA-Phonobase	8,851
Music Library of Greece of The Friends of Music Society	5,065
Bibliothèque nationale de France	4,176
Netherlands Institute for Sound and Vision	2,710

Table 1: A few of the archives that contribute to Europeana Sounds. How many unique items are in the shared data set?

3 Acoustic Fingerprinting

Some possible applications of duplicate detection have been presented in the previous section, now we see how they can be put into practice. It is clear that naively comparing every audio fragment – e.g. every five seconds – with all other audio in an archive quickly becomes impractical, especially for medium-to-large size archives. Adding robustness to speed changes to this naive approach makes it downright impossible. An efficient alternative is needed and this is where *acoustic fingerprinting techniques* comes into play, a well researched MIR topic.

The aim of acoustic fingerprinting is to generate a small representation of an audio signal that can be used to reliably identify identical, or recognise similar, audio signals in a large set of reference audio. One of the main challenges is to design a system so that the reference database can grow to contain millions of entries. Over the years several efficient acoustic fingerprinting methods have been introduced Wang (2003); Haitsma and Kalker (2002); Ellis et al. (2011); Allamanche (2001). These methods perform well, even with degraded audio quality and with industrial sized reference databases. However, these systems are not designed to handle duplicate detection when speed is changed between the original and duplicate. For this end, fingerprinting system robust against speed changes are desired.

Some fingerprinting systems have been developed that take pitch-shifts into account Fenet et al. (2011); Bellettini and Mazzini (2008); Ramona and Peeters (2013) without allowing time-scale modification. Others are designed to handle both pitch and time-scale modification Zhu et al. (2010); Malekesmaeili and Ward (2013). The system by Zhu et al. (2010) employs an image processing algorithm on an auditory image to counter time-scale modification and pitch-shifts. Unfortunately, the system is computationally expensive, it iterates the whole database to find a match. The system by Malekesmaeili and Ward (2013) allows extreme pitch-shifting and time-stretching, but has the same problem.

The ideas behind both (Six and Leman, 2014; Sonnleitner and Widmer, 2014) allow efficient duplicate detection robust to speed changes. The systems are built mainly with recognition of original tracks in DJ-sets in mind. The original tracks used in DJ-sets are manipulated in various ways and often speed is changed as well. The problem translates almost directly to duplicate detection for archives. The respective research articles show that these systems are efficient and able to recognise audio with a ±30% speed change.

Only Six and Leman (2014) seems directly applicable in practice since it is the only system for which there is downloadable software and documentation available. It can be downloaded from `http://panako.be` and has been tested with datasets containing 30,000 tracks on one dated computer. The output is data about duplicates: which items are present more than once, together with time offsets.

4 The IPEM archive as a case study

The Institute for Psychoacoustics and Electronic Music (IPEM) was founded in 1963 as a joint venture between the Belgian Radio and Television broadcasting company (BRT) and Ghent University ("Rijksuniversiteit Gent" at the time). IPEM soon established itself as production studio for electroacoustic music and artistic experimentation, and later on as documentation centre for contemporary music. Today, its archive stores audio recordings as well as photographs, videos, texts, and concert brochures, that document the international and institutional scene of the musical avant-garde in the area of Flanders and Belgium between the 1960s and 1970s. The audio archive we consider in this article comprises more than a thousand open-reel tapes with sketches, preparatory material and finished works, plus an additional 800 tapes from the collection of composer Louis De Meester.

The IPEM archive is an interesting case study for duplicate detection because it is an archive that, for several reasons, got *digitised twice* over the past fifteen years. And it is not an infrequent example. Technology evolved greatly in the past two decades, raising the standards for quality digital audio. In the first digitisation campaign, the archive was stored on Compact Discs (CDs), encoded at 44.1kHz/16bit Leman et al. (2001). Recently the archive was digitised again at today's quality standards (96kHz/24bit) and stored on Redundant Arrays of Independent Disks. When digitisation needs to be done all over again, there is little room for shortcuts: the work needs to be carried out with the best resources and methods, or the investment does not make sense. Fortunately, when it comes to meta-data, some effort can be spared. Content analysis and description is a time consuming task and requires expert staff. The first digitised IPEM archive was segmented (track recognition) and described (catalogued) in its entirety. The more recent digitised IPEM archive is currently sitting on state-of-the-art storage devices, but its content is not accessible because cataloguing staff is currently missing. Without meta-data, the potential (cultural, economical) of digital archives remains unexploited. Duplicate detection can turn this situation around by migrating the old meta-data to the new archive at a very low cost. And it can be done even between a set of segmented audio (first digitisation, tracks correspond to music pieces) and long unsegmented audio (second digitisation, tracks correspond to the entire length of each tape side).

So duplicate detection technology makes it possible to **re-use segmentation boundaries** from the first digitisation and automatically **link meta-data** to the new audio set. Error checking may still be desirable, but the amount of work for the cataloguing staff is dramatically reduced, and in the meantime the archive is already accessible (searchable). Thus, qualitative content description might be done once and for all, making it worth investing in it, because it can be transferred to subsequent audio sets and, in case, simply enriched and improved; on the other hand, technology keeps evolving and the need for future re-digitisation cannot be entirely excluded, with the tremendous cost that comes with it.

5 Deduplication in practice

In this section, the practical functioning of Panako is described. The Panako acoustic fingerprinting suite is Java software and needs a recent Java Runtime. The Java Runtime is the only dependency for the Panako system, no other software needs to be installed. Java makes the application multi-platform and compatible with most software environments. It has a command-line interface, users are expected to have a basic understanding of their command line environment.

Panako contains a deduplicate command which expects either a list of audio files or a text file that contains the full path of audio files separated by newlines. This text file approach is more practical on large archives. On a Unix systems the following two commands deduplicate an archive located in the current directory:

```
find .  -iname "*.wav > archive.txt"
java -jar panako.jar dedup archive.txt > results.txt
```

After a while, `results.txt` will contain the full path of duplicate files together with the time at which the duplicate audio was detected.

Several parameters need to be set for a successful deduplication. The main parameters determine the granularity level, allowed modifications and performance levels. The granularity level determines the size of the audio fragments that are used for deduplication. If this is set to 20 seconds instead of 10, then the number of queries is, obviously, halved. If speed is expected to be relatively stable, a parameter can be set to limit the allowed speed change. The performance can be modified by choosing the number of fingerprints that are extracted per second. The parameters determine several tradeoffs between query speed, storage size, and retrieval performance. The default parameters should have the system perform reasonably effectively in most cases.

The indirect applications such as linking meta-data needs custom glue scripts. This code takes the raw results on duplicates and according to specific meta-data standards and requirements either modifies or merges meta-data according to self-defined rules. Reuse of segmentation boundaries needs similar custom glue code. Since these tasks are very dependent on file formats, database types, meta-data formats and context in general it is hard to offer a general solutions. This means that while the duplicate detection system is relatively user friendly and ready to use, applying it still needs a software developer but not, and this is crucial, an MIR specialist.

6 Conclusions

In this paper we discussed a practical solution for duplicate detection and applications of duplicate detection. More specifically applications were discussed to complement meta-data, to link or merge digital music archives, to improve listening experiences and to re-use segmentation data.
The case study of the IPEM archive was presented: a digitised archive containing avant-garde electronic music. The aim of this paper was to make these techniques and applications better known to the community of archivists.

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.

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

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.

Haitsma, J. and Kalker, T. (2002). A highly robust audio fingerprinting system. In *Proceedings of the 3th International Symposium on Music Information Retrieval (ISMIR 2002)*.

IFLA - Audiovisual and Multimedia Section (2002). Guidelines for digitization projects: for collections and holdings in the public domain, particularly those held by libraries and archives. Technical report, International Federation of Library Associations and Institutions (IFLA), Paris (France).

Leman, M., Dierickx, J., and Martens, G. (2001). The ipem-archive conservation and digitalization project. *Journal of New Music Research*, 30(4):389–393.

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

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. 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)*, pages 1–6.

Sonnleitner, R. and Widmer, G. (2014). Quad-based Audio Fingerprinting Robust To Time And Frequency Scaling. In *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*.

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.

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.