A Large Publicly Accessible Prototype Audio Database for Music Research

Cory McKay
McGill University
Montreal, Quebec, Canada
cory.mckay@mail.mcgill.ca

Daniel McEnnis
McGill University
Montreal, Quebec, Canada
daniel.mcennis@mail.mcgill.ca

Ichiro Fujinaga
McGill University
Montreal, Quebec, Canada
ich@music.mcgill.ca

Abstract

This paper introduces Codaich, a large and diverse publicly accessible database of musical recordings for use in music information retrieval (MIR) research. The issues that must be dealt with when constructing such a database are discussed, as are ways of addressing these problems. It is suggested that copyright restrictions may be overcome by allowing users to make customized feature extraction queries rather than allowing direct access to recordings themselves. The jMusicMetaManager software is introduced as a tool for improving metadata associated with recordings by automatically detecting inconsistencies and redundancies.

Keywords: Music database, MP3s, features, metadata.

1. Introduction

The maturation of the MIR field is increasingly requiring researchers to move beyond work involving simplistic musical datasets to more expansive studies that require much larger, more varied and carefully annotated (i.e., labelled with metadata) collections. This is necessary in order to develop and validate MIR tools that can be applied to the vast and varied universe of music that exists outside of the lab. Developing a framework for building and maintaining large music databases is also becoming increasingly important as libraries digitize their collections and on-line commercial databases continue to grow. In addition, high quality datasets are needed to provide ground truth that can be used to compare the performance of different systems in competitions such as MIREX [1]. Unfortunately, problems relating to copyright laws and poor metadata annotations have hindered efforts by researchers to form, share and use high quality collections.

This paper discusses approaches to overcoming these limitations and presents a prototype database of recordings with these ideas in mind. In particular, the following list is proposed as a basic set of requirements that should be met by any music database that is to effectively meet the broad needs of current MIR research:

- Data should be freely and legally distributable to researchers.
- The database should contain many different types of music.
- The database should include many thousands of recordings. This is important not only to allow sufficient variety, but also to avoid research overuse of a relatively small number of recordings, which can result in overtraining. Furthermore, even good quality annotations will inevitably contain some errors, and a large database helps to average out such noise.
- The database should include a significant amount of commercial music, although independent music can certainly play a role as well. The vast majority of end users are interested primarily in professionally produced music, so MIR systems must demonstrate that they are able to deal with such music.
- Each recording should be annotated with as diverse a range of metadata fields as possible in order to make the database usable as ground truth for as wide a range of research as possible.
- It should ideally be possible to label segments of recordings as well as recordings as a whole.
- Annotations of subjective fields such as genre or mood should include a wide range of candidate categories, as simply allowing ten or so coarse categories is unrealistic.
- Annotations should be correct, complete and consistent.
- It should be possible to assign multiple independent values to a single field so that, for example, a recording could be classified as both swing and blues.
- The ability to construct ontological structures between fields could be useful.
- Metadata should be made available to users in formats that are easy to both manually browse and automatically parse.
- Automatic tools should be available to validate metadata and generate profiles of the database.
- Entire recordings should be accessible, at least indirectly, not just short excerpts. Each researcher should be able to choose how much and what parts of recordings are to be studied.
- Given that different compression methods can influence extracted feature values, the audio format(s) most commonly used by the public should be adopted in order to reflect realistic conditions. This is important for many types of end user oriented research, although uncompressed audio is also useful for some theoretical research. A variety of encoders and bit rates should be used for similar reasons.
- It should be easy to add new recordings and their metadata.

2. Existing Databases

MIR researchers have traditionally assembled their own musical datasets for training and/or testing, an understandable approach given the lack of alternatives. This approach is ultimately flawed, however, as collections assembled in

this manner tend to be restricted in size and variety, and can also be biased due to the characteristics of the system being developed and the musical background of the individuals assembling the database. Furthermore, such databases cannot be legally distributed for refinement, expansion and comparison if they contain copyrighted material. Although such datasets have sometimes been "unofficially" distributed among researchers in the past, this quasi-legal sharing is not viable in the long term.

An alternative approach is to make use of recordings that are in the public domain, such as through sites like GarageBand [2]. Homburg et al. [3] have explored such an approach, and have also incorporated multiple views and high quality annotations. Unfortunately, this approach generally restricts one to a limited number of recordings that are usually either very old, are by amateur musicians or consist of only short low quality extracts. These limitations are demonstrated by Homburg et al.'s database, which, despite its many advantages, contains only 1886 files, of which only 10-second segments are available.

Some improvement can be achieved by utilizing music protected under more limiting but still fairly lenient frameworks such as Creative Commons. Web sites such as Magnatune [4] and Epitonic [5], for example, allow one to preview recordings for free, and entire music collections can sometimes be licensed for research purposes at little or no cost. Unfortunately, size is still an issue. For example, at press time Magnatune had only 210 artists and 5662 songs. Also, the music available on such sites usually excludes artists contracted to major record labels, which is the music that most users are interested in. In addition, the metadata contained in ID3 tags is not entirely reliable, with problems such as multiple spellings of a genre names or strange treatment of special characters. Finally, there can still be legal restrictions on distributing downloaded music to other researchers, even if one is permitted to access and store entire recordings for free oneself.

An additional possibility is to contract arrangers and musicians to produce original recordings for use in research, as was done in constructing the RWC database [6]. Although this approach has the important advantage that it overcomes copyright protection, costs prevent it from scaling to any reasonably large number of recordings. There can also be doubts as to how well such original music simulates what one encounters in the real world.

3. Overcoming Copyright Limitations

The copyright laws of many countries make it difficult to acquire access to a sufficient range of music to meet the needs of MIR research. This puts pressure on each researcher to independently collect and annotate his or her own dataset, a process that, in addition to the problems discussed in Section 2, also wastes significant amounts of time in duplicated effort.

The motivation behind copyright laws is not to hinder research, of course, but to protect intellectual property and prevent pirating. Discussions with legal experts at past ISMIR conferences seem to indicate that there is likely no legal obstacle to the distribution of information extracted from recordings, as long as such information cannot be used to reconstruct the music itself. This means that many of the features that are typically extracted form music in MIR research can in fact be publicly distributed, even when the music itself cannot. Even features such as MFCCs can only reconstruct a very poor facsimile of the original music under certain parameterizations.

Since features and metadata are essentially what many MIR researchers are interested in, publicly distributing this data rather than the music itself is a useful way of circumventing copyright limitations. Features extracted from legal local music collections can thus be combined into databases so that researchers can effectively share their music.

Simply extracting and publicly posting stock features is insufficient, however, as an important part of MIR research involves developing new and specialized features. Furthermore, different researchers will want features extracted with different parameters (e.g., window size and overlap, downsampling, amplitude normalization, etc.). The jAudio software [7], an open-source Java-based feature extraction package, presents a solution to these problems by allowing specification of extraction parameters and by making it a relatively simple matter for researchers to design their own features and add them to the jAudio framework.

Part of the On-demand Metadata Extraction Network (OMEN) project [8] has involved integrating OMEN services with jAudio so that it can receive and execute parameterized feature extraction requests as well as deploy new feature classes while it is running on a network. This means that a database of musical recordings can be stored privately on one or more servers, which can receive and process customized feature extraction requests. Researchers can thus access needed features without ever violating copyright laws by accessing the music itself. The only intervention needed at the server is examination of requests to ensure that they will not provide sufficient information to reconstruct the music.

A good way to provide the computing power needed to service feature extraction requests is to make use of distributed computing. Libraries offer a particularly suitable resource in this respect, as they have large networks of computers around the world that go unused outside of opening hours. They also often have large music collections from which features could be automatically extracted and made publicly available with jAudio and OMEN.

4. Achieving High Quality Annotations

There can be difficulty involved in properly annotating the metadata associated with music recordings. Both significant amounts of time and extensive musicological knowl-

edge are often needed, both of which can be lacking when performing technical MIR research where dataset collection is only part of a larger project. Accurate and consistent annotation is nonetheless essential, as the metadata found in annotations will often serve as the ground truth for training and evaluating systems.

Subjective fields such as genre or mood are particularly problematic, as multiple interpretations may be valid. Such fields are nonetheless important when constructing a database, as many researchers and end users are interested in them. Extra care must be taken when annotating such fields, in terms of both which candidate categories are used and how they are assigned. A number of associated issues are well-discussed elsewhere [9].

It is necessary to consider alternatives to manual metadata entry when constructing a database, as this would be very time consuming for large databases. This is typically done by extracting metadata from the ID3 tags of MP3s or from metadata management services such as Gracenote CDDB [10]. Unfortunately, although these sources do save significant amounts of time, they tend to be at best noisy and inconsistent, and at worst entirely incorrect.

It is necessary to correct the metadata derived from such sources if they are to be used. The Java-based jMusicMetaManager software was therefore built to automatically error check the metadata of music databases.

One of the important problems that jMusicMetaManager deals with are the inconsistencies and redundancies caused by multiple spellings that are often found for entries that should be identical. For example, uncorrected occurrences of both "Lynyrd Skynyrd" and "Leonard Skinard," or of the multiple valid spellings of composers such as Stravinsky, would be problematic for an artist identification system that would perceive them as different artists.

At its simplest level, jMusicMetaManager calculates the case-insensitive Levenshtein (edit) distance between each pair of entries for a given field. A threshold is then used to determine whether two entries are likely to in fact correspond to the same true value. This threshold is dynamically weighted by the length of the strings and whether their other fields are similar. This is done separately once each for the artist, composer, title, album and genre fields. In the case of titles, recording length is also considered, as two recordings might correctly have the same title but be performed entirely differently (e.g., an original Led Zeppelin song compared with a Dread Zeppelin cover, or live and studio versions of the same song both by Led Zeppelin).

This approach, while helpful, is too simplistic to detect the full range of problems that one finds in practice. Additional processing was therefore implemented and additional post-modification distances were calculated. For example:

- Instances of "The " were removed (e.g., "The Police" should match "Police").
- Occurrences of " and " were replaced with " & " (e.g., "Simon and Garfunkel" should match "Simon & Garfunkel").
- Personal titles were converted to abbreviations (e.g., "Doctor John" should mach "Dr. John").
- Instances of "in'" were replaced with "ing" (e.g., "Breakin' Down" should match "Breaking Down").
- Punctuation and brackets were removed (e.g., "REM" should match "R.E.M.").
- Spaces were removed, as their omission is a common typo.
- Numbers were removed from the beginnings of titles, as track numbers are sometimes encoded in titles.
- Word orders were rearranged (e.g., "Ella Fitzgerald" should match "Fitzgerald, Ella," and "Django Reinhardt & Stéphane Grappelli" should match "Stéphane Grappelli & Django Reinhardt").

jMusicMetaManager also automatically generates a variety of HTML-formatted statistical reports about music collections, including multiple data summary views and breakdowns of co-occurrences between artists, composers, albums and genres. This allows one to easily acquire and publish HTML database profiles.

Users often need a graphical interface for viewing and editing a database's metadata. It was therefore decided to link jMusicMetaManager to the Apple iTunes software, which is not only free, well-designed, and commonly used, but also includes an easily parsed XML-based file format. iTunes, in addition, has the important advantage that it saves metadata modifications directly to the ID3 tags of MP3s as well as to its own files, which means that the recordings can easily be disassociated from iTunes if needed. iTunes can also access Gracenote's metadata automatically, which can then be cleaned with jMusicMetaManager.

jMusicMetaManager can extract metadata from iTunes XML files as well as directly from MP3 ID3 tags. Since MIR systems do not typically read these formats, jMusicMetaManager can also be used to generate ground-truth data formatted in ACE XML [11] or Weka ARFF [12] formats. This is also important because iTunes XML has a limited number of fields, only allows one genre per recording, does not allow ontological structuring and does not allow segmented annotations of recordings. More flexible file formats such as ACE XML allow access to expanded expressiveness when required.

5. Details of the Codaich Database

A prototype database named Codaich (Gaelic for "share") was constructed using the majority of the guidelines and tools discussed in this paper. It currently consists of 20,849 MP3 recordings from 1941 artists. Details on the database may be access via iTunes XML, ACE XML, Weka ARFF or jMusicMetadata HTML files. Codaich is intended to be integrated with OMEN so that features values may be publicly accessed.

Metadata fields (including Title, Performer, Composer, Album, Track Number, Disc Number, Year, Genre, Bit Rate and Track Duration) were originally extracted from Gracenote CDDB and pre-existing ID3 tags. These were

then cleaned using the jMusicMetaManager software, and final manual improvements were made with iTunes when necessary, in consultation with the AllMusic Guide [13]. The recordings are annotated using a total of 53 candidate genres, which are distributed among the coarse categories of popular, world, classical and jazz. Efforts were made to achieve as stylistically diverse a collection as possible.

The MP3 audio format was chosen because it is by far the most popular format. Some MP3s were ripped from CDs, some were recorded from tapes and LPs and some were found as pre-encoded MP3s. A variety of encoders and bit rates were used. This diversity simulates what one finds in the real world.

The recordings were acquired from the Marvin Duchow Music Library, from contributions from the personal collections of members of the McGill Music Technology Area and from the in-house database of Douglas Eck's lab at the Université de Montréal.

The jMusicMetaManager software and information on accessing the Codaich database may be found at http://sourceforge.net/projects/jmir.

6. Conclusions and Future Research

This paper has emphasized the importance of building large publicly accessible databases of musical recordings for use in MIR research, and has discussed important issues to consider when constructing such a database. The Codaich prototype database was presented, and the OMEN framework was suggested as a means of distributing features extracted from this database without violating copyright laws. The jMusicMetaManager software was introduced as a tool for generating profiles of music collections, as well as for cleaning recordings' metadata by detecting inconsistencies and redundancies.

The Codaich database is still growing rapidly, and the authors will continue to add many more recordings to it. It is hoped that in the future other researchers will also contribute their collections using the OMEN framework.

A priority for future research is the development of improved data mining software to automatically mine multiple sources on the web to improve entries for fields such as genre. Additional fields such as mood will also be added in order to make the annotations suitable for a wider range of MIR research, and functionality for incorporating sophisticated ontologies and multiple entries per field will be developed. In addition, there are plans to integrate audio fingerprinting technology into jMusicMetaManager to fill in missing fields and detect incorrect labels. Also, automated name authority control using a standardized reference such as the U.S. Library of Congress will be incorporated [14]. Finally, the framework discussed here will be adapted to construct a database of symbolic recordings in formats such as MIDI and Humdrum kern.

7. Acknowledgments

We would like to thank the helpful and patient librarians at the Marvin Duchow Music Library for accommodating our many requests for CDs. We would also like to thank Douglas Eck for giving us access to his lab's database as well as the other individuals who contributed recordings. Finally, we are grateful to the Social Sciences and Humanities Research Council of Canada and the Canada Foundation for Innovation for their financial support.

References

[1] J. S. Downie, K. West, A. Ehman, and E. Vincent, "The 2005 Music Information retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview," in *Proceedings of the International Computer Music Conference*, 2005, pp. 320–3.

[2] "GarageBand.com," [Web site] 2006, [2006 March 27], Available: http://www.garageband.com

[3] H. Homburg, I. Mierswa, B. Moller, K. Morik, and M. Wurst, "A benchmark dataset for audio classification and clustering," in *Proceedings of the International Conference on Music Information Retrieval,* 2005, pp. 528–31.

[4] "Magnatune: MP3 music and music licensing," [Web site] 2006, [2006 March 27], Available: http://magnatune.com

[5] "Epitonic.com: Hi quality free and legal MP3 music," [Web site] 2006, [2006 March 27], Available: http://www.epitonic.com

[6] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *Proceedings of the International Conference on Music Information Retrieval*, 2002, pp. 287–8.

[7] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle, "jAudio: A feature extraction library," in *Proceedings of the International Conference on Music Information Retrieval,* 2005, pp. 600–3.

[8] I. Fujinaga, and D. McEnnis. "On-demand Metadata Extraction Network," in *Proceedings of the Joint Conference on Digital Libraries*, 2006.

[9] J. J. Aucouturier, and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 1–12, 2003.

[10] "Gracenote," [Web site] 2006, [2006 March 29], Available: http://www.gracenote.com

[11] C. McKay, D. McEnnis, R. Fiebrink, and I. Fujinaga. "ACE: A general-purpose classification ensemble optimization framework," in *Proceedings of the International Computer Music Conference*, 2005, pp. 161–4.

[12] I. H. Witten, and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques,* 2nd ed., New York: Morgan Kaufman, 2005.

[13] "AllMusic Guide," [Web site] 2006, [2006 March 29], Available: http://www.allmusic.com

[14] T. DiLauro, G. S. Choudhury, M. Patton, and J. W. Warner, "Automated Name Authority Control and Enhanced Searching," *D-Lib Magazine,* vol. 7, no. 4, 2001.