AUDIO ISSUES IN MIR EVALUATION

Josh Reiss *Mark Sandler*
Centre for Digital Music
Queen Mary, University of London
Mile End Road, London E14NS
United Kingdom

ABSTRACT

Several projects are underway to create music testbeds to suit the needs of the music analysis and music information retrieval (MIR) communities. Furthermore, there are plans to unify the testbeds into a grid whereby research can be performed in a distributed manner. Thus the issue of audio file formats has come to the forefront. The creators of a music library or MIR testbed are confronted with a variety of questions pertaining to file formats, their quality, metadata, and copyright issues. We discuss the various formats, their advantages and disadvantages, and give a set of guidelines and recommendations. This document is a positional paper. It is intended to foster discussion and not as a definitive statement. Nevertheless, it is hoped that the proposals put forth here may serve as a guideline to use in construction of an MIR evaluation testbed.

1. OVERVIEW OF AUDIO FORMATS

There exists a bewildering variety of audio formats. Distinguishing between them is best done via their intended purpose. Mp3, for example, is designed for compressed delivery of audio, whereas Broadcast WAV is intended for uncompressed storage and exchange of production quality audio.. In order to differentiate and classify the audio formats available, we distinguish between raw audio formats, compressed formats, and multimedia interchange wrappers. This classification is generalized and there are many exceptions. For instance, WAV format supports compression, and AES31 describes file formats, compression schemes and wrappers. However, the following groups the formats based on their complexity, general usage and purpose, and thus serves as a good guide to the options available for audio formats used in a testbed.

1.1. Raw Audio Formats

The most common audio formats for end users are based on simple, open standards that have been designed and promoted by companies for certain platforms. These include Apple's AIFF format for the Mac, Sun's au format for UNIX, and the WAV format for Windows, developed by Microsoft and IBM. Despite this legacy, these formats can be played on almost any computer using many audio applications, and contain no features specific only to their original intended platforms.

These 3 formats share many features. They all (typically) are intended for storing uncompressed, PCM-encoded, raw audio in a single binary file. They each support a variety of bit rates, sample rates and channels, and contain a header containing such information. They differ primarily in the byte order (little endian or big endian), and how the header is constructed. Of these formats, WAV and AIFF are by far the most common. Almost all audio workstations support both.

1.2. Broadcast WAV

Broadcast WAV is highly relevant because it is used for the creation of master recordings. It represents the highest quality audio format available, and in the richest form. The European Broadcast Union (EBU) introduced the format to allow file exchange between digital audio workstations (DAWs) used in radio and television production[1]. It is now the standard for file storage and exchange in the audio production industry. This implies that almost all master recordings, including those from small studios, live recordings and remasterings of older recordings are created using Broadcast WAVs. Even specialized workstations using proprietary software allow for import and export in this format.

All WAV file players should be able to recognize and play Broadcast WAV. The Broadcast WAV format is similar to a WAV file except it contains an additional header which provides information about the originator, a time stamp, and sound sequence description metadata. The basic audio format is 16-bit linear PCM sampled at 48kHz, but additional sample rates and bit depths may also be used, and MPEG-encoded audio is supported.

Broadcast WAV files, as used in the mastering and editing process, are often stored as multiple mono files. A multi-track recording may thus contain a large number of high quality files and an edit decision list is needed to describe how they are combined in the final mix.

1.3. Compressed Audio Formats

The choices of compressed audio formats are almost endless. The problem has arisen since many standards bodies and many companies have released different compressed formats, and these have all found niches where they have become popular and entrenched. However, for the purposes of a testbed, only the most relevant ones will be considered. Here, relevance may be defined in terms of quality, popularity and ease-of-use.

1.3.1. Lossless Compression Schemes

Few lossless compression schemes have seen widespread use with audio. This is because the compression achieved by lossless means is usually insufficient to warrant the added complexity. Lossless compression schemes do not seem appropriate for use in

a testbed. Although they provide no loss in quality, many players would not support them, they require additional computational cost in encoding and decoding, and are not typically used for many of the audio formats under consideration. But the overriding reason why they are not necessary is that if it can be assumed that there is ample storage space for the testbed, then the primary purpose for lossless compression of audio is irrelevant. Conversely, if storage capacity is significantly limited, then only lossy compression schemes provide enough reduction in file size to warrant their use.

1.3.2. Lossy Compression Schemes

The issues concerning lossy compression schemes for encoding of audio files in a testbed are far more pertinent. Lossy schemes have widespread use, and thus music analysis and processing algorithms must be robust against their use. Conversely, they offer significant degradation of quality and hence ensure that the audio becomes further removed from the groundtruth. A full discussion of the pros and cons of lossy compression in MIR systems is presented in Section 2, along with recommendations concerning their use. In this section, we discuss the options available for lossy compression.

Different lossy compression schemes, or codecs, have seen acceptance for use in different settings. Audio encoded for network (internet) transmission, for use on DVDs and in theatres, for use with telephony and for digital broadcasting, have all seen the acceptance of various compression schemes. These schemes include both proprietary methods proposed by companies and open methods from international standards bodies and from industrial consortiums. Although open compression methods are obviously preferred in research settings, many audio collections that might be donated to a testbed could be encoded using proprietary methods, and these methods are popular enough to be considered as tests of robustness. Thus they will also be discussed.

Dolby Labs has been a pioneer in the development of codecs. Dolby Digital (AC-3) provides multichannel surround sound in cinemas and on DVDs, and digital broadcast television, cable, and satellite systems. It enables the transmission and storage of up to five full-range audio channels, plus a low-frequency effects channel in less space than is required for one linear PCM-coded channel. AC-3 uses perceptual models to throw away audio in a signal which the listener is unlikely to hear, and multiplexing to fold 6 channels of sound (5.1) down to 2 channels of data, which fit more easily on DVD media. Similar multichannel high fidelity compression schemes include MPEG-2, DTS and SDDS.

Codecs based on perceptual models gained widespread acceptance with the introduction of mp3. Mp3 compression, like AC-3, uses information about human hearing to make decisions about which parts of the sound data are extraneous. These extraneous frequency components are coarsely quantised with virtually no discernible effect on the perception of the sounds that result when the data is converted to analog and played. Its large compression rates and open standard has lead to mp3 audio becoming the preferred format for audio distribution online.

Many perceptual coders have better performance than mp3 and have achieved widespread acceptance. Microsoft's WMA, for instance, encodes audio at equivalent quality to mp3 with only half the size. It has widespread support due to Microsoft Windows' large user base, but has many Windows-specific features. The Advanced Audio Coding (AAC) developed and standardized by AT&T, Dolby, Fraunhofer, and Sony, is a high-quality alternative. Coding efficiency is similar to that of WMA. AAC is supported by many manufacturers as the successor to mp3.

From the descriptions above, it is clear that there are a significant number of important lossy formats. Despite mp3's popularity for transmission of audio over the internet, large music libraries, especially from commercial enterprises, will often use alternative compression schemes. The choice of compression method, if used at all, is discussed in a later section. For now, we note that, if the MIR community is soliciting for donations of copyrighted music, it is not sufficient to simply demand that all audio be encoded in mp3 format.

1.4. Exchange Formats and Wrappers

In general, the formats described above for compressed and uncompressed audio only allow for encoding of a small amount of metadata. The audio production, information retrieval, broadcasting, internet software and copyright law communities have all sought better ways to exchange the audio along with relevant metadata. For audio editing and mixing, metadata primarily involves edit decision lists. For information retrieval, metadata necessary for text-based searches by composer, performer, year, etc… must be incorporated. Audio transmission over the internet requires advanced coders for streaming as well as custom metadata attributes for advanced applications. And encoded copyright information is required under all situations in order to guarantee protections for creators and license holders.

Thus a variety of formats have been proposed for the exchange of the raw of compressed audio files. These formats typically use a *wrapper*, which contains the metadata, along with audio and other multimedia data in any of a variety of formats.

Perhaps the most popular open format to come from an individual company is Tascam's OpenTL (Open Track List) format. Although OpenTL has widespread support, especially in DAWs, it is likely to be surpassed by AAF and/or AES31. Yet it is worth mentioning because many master recordings are stored in this format, and because, like most audio wrapper and exchange formats, it supports broadcast WAV.

Interchange of production-level audio found a partial solution in OMFI (Open Media Framework Interchange). OMFI provides an open digital-media interchange format

between applications and across platforms, and supporting a variety of media. With regard to audio, OMFI allows a project, including individual tracks and editing, mixing and processing instructions, to be interchanged between different DAWs. Issues concerning suitability and dependability have lead to the introduction of Advanced Authoring Format (AAF). AAF has similar goals to OMFI, but incorporates more metadata capabilities, Microsoft's open container format Structured Storage, the management of Pluggable Effects and Codecs, and has broad industry support. AAF extensible and royalty-free, and supports digital rights management, links, and interactive content.

Whereas AAF is intended for multimedia exchange during the authoring and postproduction phases, the Material eXchange Format (MXF) is designed to facilitate transfer of finished content between broadcast systems. It has the ability to read metadata regardless of internal data format. MXF is derived from the AAF data model and the two formats are complementary. Both formats can stand on their own. A broadcast system may use only MXF and a postproduction house, just AAF, but a broadcaster with a post facility may well use both

AAF, MXF and OMFI are intended for exchange of audio, video and other media, and were not designed specifically for audio and music. The only open exchange format designed by and for the audio community is AES31.

The Audio Engineering Society Standards Committee Working Group on Audio-File Transfer and Exchange was established in response to demand for a project interchange format that may be used as a simple alternative to OMF and proprietary formats. The result is a viable technique for transferring sound files and project data. AES31 provides a set of technical specifications that, allow digital audio media and metadata to be transferred between workstations. AES31 is actually four documents. Collectively, they describe all components of an audio project exchange system.

- AES31-1 is concerned with **physical data transport**, how files can be moved from one system to another by removable media or high-speed network. AES31-1 specifies a transport compatible with FAT32 structures.
- AES31-2 focuses on **audio file format**, how the data should be arranged on the removable media or packaged for network transfer. It specifically recommends the use of Broadcast Wave files for storage of individual audio tracks.
- AES31-3 describes a **simple project structure**, using an Audio Decision List, or ADL. The ADL was modeled on conventional Edit Decision Lists, but with sample-accurate precision and parameters for multiple audio channels, crossfades, level automation and so on.
- AES31-4 is an **object-oriented project structure**. It is an extensible object model capable of describing a wide range of parameters for advanced applications.

AES31-1 through 3 have been ratified as standards[2]. AES31-4 is in the development stages. It could be based on the AAF format, and currently there is liasing between the AES standards body and AAF consortium to harmonise the two. Thus, it appears that there will soon be an internationally recognised standard for the interchange of complete audio projects, based on existing informal standards.

1.4.1. Metadata for Information Retrieval

The field of wrappers and exchange formats for audio becomes far more complex when one includes the multitude of standards proposed within the Information Retrieval and Library Science communities[3]. In general, these are not specific to raw audio, or to audio projects and master recordings. They are metadata formats, which describe how documents should be linked, and metadata such as composer, performer, year etc., should be incorporated for text searches and copyright information. These formats, and the issue of whether they should be used, is virtually unaffected by the choice of audio format, even when master recordings or an audio project is used. They are mentioned in order to understand the full range of options available and whether their use in any way conflicts affects the choice of audio format.

CIDOC, for instance, provides a set of guidelines for metadata in museum collections. The Open Archival Information System (OAIS) is a framework for preserving access to digital information over the long term. METS, the Metadata Encoding and Transmission Standard, provides XML schema for encoding hub documents for digital material. METS provides a vocabulary and syntax for identifying components that together comprise digital objects, for specifying component locations, and for expressing structural relationships. MARC is the Machine-Readable Cataloging format used to describe bibliographic materials. It facilitates cooperative cataloging and data exchange in bibliographic information systems. It is a mature, well accepted standardized set of metadata, and operates as a protocol for communication of bibliographic data. Finally (although by no means the final metadata standard), the Dublin Core Metadata Initiative promotes interoperable metadata standards and develops specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems.

To resolve issues that arise concerning audio formats, it is not necessary to know all the metadata standards and how they operate. There exist appropriate schemes to associate disparate audio files, to classify and retrieve audio in digital libraries, and to add additional metadata suitable for audio present in a music digital library (MDL) or testbed. Furthermore, these schemes are distinct from wrappers such as AES31 and AAF, which are more appropriate for storing and exchange of audio projects. Wrappers such as AES31, or even Windows Media Format 9, may be necessary if one wishes to incorporate audio as close as possible to the groundtruth, e.g., the recording masters.

1.5. Choice of Audio Format for an MIR Testbed

The preceding section indicated the range of choices that are available for the formats of audio files in a music retrieval testbed. It is clear, that most metadata formats are independent of the audio format. However, pivotal questions still remain regarding the preferred choice of audio format, especially concerning compressed and proprietary formats.

In the following sections, we list a set of guidelines that should be followed in the choice of file formats for the audio in a MIR testbed. These suggestions represent the ideals for choice of format, use of compression, use of standards, and file access and editing. Actual choices made in the creation of a testbed are limited by the files to which the creators receive access. Thus, these guidelines also serve as a list of requests for files provided for use in the testbed by copyright holders.

2. QUALITY GUIDELINES

Audio files should be presented in the highest quality format possible, ideally the original master recordings. If a compressed format is used, it should be used in tandem with the original format.

Although this may seem obvious, many people have argued against this for several reasons. Uncompressed high quality audio occupies a tremendous amount of space, whereas compressed audio can be less than one tenth the size yet still of acceptable quality for many purposes. Also, low quality formats are tremendously popular, and hence retrieval should focus on those formats. And related to the popularity of low-quality formats, retrieval and analysis methods should be robust against compression schemes and thus the compressed audio should be used in order to guarantee the robustness of any algorithm. However, as shall be explained, the use of compressed or low-quality formats severely limits the quality of retrieval as well as the scope of analysis tasks that are possible. Furthermore, the benefits of compressed audio, with the exception of small file size, can be achieved more effectively and simply if the original audio is stored.

2.1. Compression as Error

The most accurate retrieval can be achieved using the highest quality audio. Any lossy compression represents a distortion of the signal. The groundtruth, which represents the actual original signal(s) without errors introduced by acquisition, processing or compression, yields the most information which can be used to aid retrieval. Furthermore, compression often involves additional processing of the signal whereby distortions are introduced. These processing steps involve considerable distortion to the signal. They may introduce artifacts and hinder the retrieval of relevant documents.

2.1.1. Preprocessing

Often overlooked is the preprocessing that occurs before compression. Before a signal is compressed, there is usually a preparation stage whereby it is companded (to modify the volume range), equalized (to normalise the strength over the frequency range) and/or boosted (increased strength at certain frequencies). Each processing method modifies the signal and makes it further removed from the original. In addition it is often cleaned, which may remove important musical components as well as background noise. These processing steps all combine to make MIR more difficult. And since it is usually not known exactly what processing was performed, it becomes difficult to differentiate between retrieval failure (low precision or low recall) due to problems with the retrieval method or due to excessive processing on the audio files.

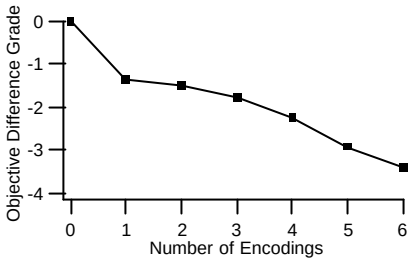

Figure 1. The effect of multiple encodings on perceptual audio quality. The Objective Difference Grade measures perceived difference in quality between a reference and test signal.

2.1.2. Repeated Encoding

This phenomenon can occur regardless of whether the same or different coders are used. In perceptual codecs, the spectral representation of the signal is altered slightly by the quantization process. The lower the bitrate, the coarser the quantization that is required to represent the signal. In this way, distortion is introduced which can be modeled as additive coding noise. This noise is shaped according to criteria estimated by the perceptual model. Figure 1 demonstrates this phenomenon where multiple encodings are used to convert a 2 channel, 48khz, 16 bit, 57 second sample of music to 128kbps mp3 format. The Objective Difference Grade[4] is a perceptual audio quality measure, based on the ITU-R recommendation BS.1387, which rates the difference between two signals on a scale from 0 (imperceptible) to -4 (very annoying).

With increasing deployment of low bitrate audio coding, use of audio compression can happen at various stages between initial performance and final playback, or deployment in an audio testbed. Audio processing and transmission operations, change of audio coding formats and/or bitrates, may combine to create multiple cycles of decoding, processing and encoding of audio content. Quantization noise from each cycle accumulates and leads to a progressive drop in audio quality. The resultant distortion quickly becomes audible and more and more problematic with each generation.

2.2. Introduction of Artifacts

A dangerous aspect of compression is that it can introduce artifacts. Not only are the inaudible parts of the

signal affected, and the audible parts carry less information, but the audible signal may become drastically modified. In effect, this guarantees that even a robust similarity measure may fail if artifacts have been introduced. And furthermore, the inaudible artifacts may still affect the reliability of analysis algorithms.

2.2.1. Pre-echo

This is the result of using a large block size in processing transient signals. When a transient occurs, a perceptual model will allocate only a few bits to each of the quantizers in the subbands because a transient signal in the time domain will spread out in frequency over many subbands**[5]**. When the compressed data is decoded, the subbands samples are reconstructed and the quantization noise, which was supposed to be fully masked, may now spread over the entire block. Therefore, this noise will also precede the time domain transient. The quantization noise announces the transient in advance, producing a potentially audible artifact. It can be noticed prior to the signal attack as a pre-echo.

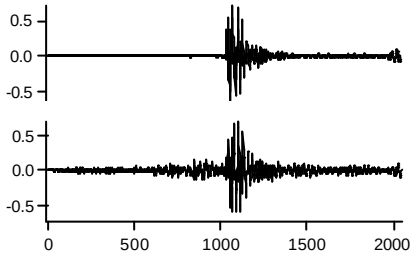

Figure 2. Original signal(top) and pre-echoes(bottom) from castanets with block processing of 2048 samples.

2.2.2. Aliasing (low frequency sampling)

Aliasing is well-understood but often overlooked in the coding process. If too low a sampling rate is used, the signal can impersonate another signal at lower frequency. Since many signals are sampled at very close to Nyquist, design of suitable anti-aliasing filters is difficult.

Aliasing introduces additional problems when used in conjunction with compression. Aliasing results in the quantization noise introduced into a specific subband creating additional noise at different frequency locations. Thus, frequency components that have negligible effect on audio quality become non-negligible when they are aliased down into more audible frequencies. Although there are many ways that aliasing problems can be avoided, it is not guaranteed that all popular audio coders will have implemented these methods.

2.2.3. Birdies (masking)

The most used perceptual measure in audio coding is the masking threshold. For low bit rates, slight variations of the masked threshold from frame to frame leads to very different bit assignments. As a result, some groups of spectral coefficients may appear and disappear. This spurious energy constitutes several auditory objects, which are different from the main one and thus clearly perceived. These artifacts, known as birdies, have been reported both for the tuning of audio codecs and for objective perceptual assessment methods.

2.2.4. Loss of Stereo Image

Directional localization of sound depends on evaluation of spatial cues by the human auditory system. So the fidelity of the stereo image depends on the coder's ability to preserve critical cues. Intensity stereo coding exploits the fact that perception of high frequency sound relies mostly on the envelope rather than on the waveform itself. Thus, it may be sufficient to code the envelope of such a signal instead of its waveform. This is done by transmitting a shared set of spectral coefficients (carrier signal) instead of separate sets for each channel. In the decoder, the carrier signal is scaled independently for each signal channel to match its original average envelope for the respective coder frame. The scaling information is calculated and transmitted once for each group of spectral coefficients. As a consequence, all signals reconstructed from a single carrier are scaled versions of each other. They have the same envelope structure for the duration of a frame. For transient signals with dissimilar envelopes in different channels, the original distribution of onsets between coded channels cannot be recovered. In a stereophonic recording of an applauding audience, envelopes will be very different in the right and left channel due to the distinct clapping events happening at different times in both channels.

After the intensity stereo encoding / decoding process, there is cross-talk between the channels. The fine time structure of the signals is mostly the same in both channels. Perceptually important signal onsets propagate to the opposite channel. This results in significant loss of stereo image quality. The spatial impression tends to narrow and the perceived stereo image collapses into the center position. For signals with uncorrelated transient information in each channel, like an applauding audience, the signal may seem to disappear from different locations at different times.

2.3. Quality Requirements for Audio Analysis

Compression hinders the ability to analyse audio files correctly. Each lossy compression method uses an algorithm that analyses the signal and determines which components can be removed without serious degradation of quality. For instance, the WMA Codec is based on the Malvar wavelet. If this wavelet is used as a similarity measure in an MIR system, then the results will be biased towards ranking any WMA compressed files as highly similar. Similarly, music processing methods which encode data in the masked frequencies, or similarity measures which incorporate high frequency comparisons, may fail on mp3s because masking and high frequency removal are integral to mp3 encoding. This has disastrous consequences for instrument templates, since mp3 compressed audio is not an accurate representation of the frequency content (including harmonics) when a note is played on an

instrument. The creation of instrument templates is often a useful component of an MIR system.

The production of inaudible artifacts due to compression is also problematic. Even inaudible artifacts may still hinder analysis algorithms. Both pre-echo and sampling rate reduction, for instance, increase the uncertainty of the time at which an event occurs. Thus accurate measurement of note onsets becomes more difficult.

2.4. Why use the master recordings?

The previous section provides a variety of reasons why a compressed or low-quality audio signal should not be used as the preferred audio format for an MIR testbed. Although this justifies the use of high quality audio, it does not make the case for the use of original master recordings. Master recordings are very large, and may prove difficult to acquire. However, their use is justified due to their quality, richness, metadata and the fact that they provide capabilities for far more analysis, processing, and investigation of retrieval methods.

2.4.1. Highest quality

The master recordings guarantee the highest quality digital recording of a performance that is available. As such, they represent the closest to the groundtruth that may be achieved. They provide a wealth of information far in excess of a CD quality recording. It is not uncommon for master recordings to incorporate, for instance, 16 24-bit tracks, sampled at 96kHz or higher. Since it has already been established that processing and compression introduce errors which may be detrimental to the evaluation of MIR methods, the use of master recordings guarantees the least chance of these problems occurring. High quality allows for more accurate transcriptions, measurements of harmonic contours, instrument recognition, and so on. All of which are tools which may be applied to an MIR system. Hence MIR systems can perform more accurately when they have access to the master recordings in the corpus.

2.4.2. Tests Robustness

An argument given in favour of low-quality or compressed audio is that good MIR systems should be robust to the distortions produced by compression. Since mp3 is by far the most popular format for internet delivery of audio, it makes sense for audio files to be compressed and encoded in mp3 format. However, limiting the audio to any one compressed format restricts the ability to test for an algorithm's robustness against other formats. The best way to test for robustness would be to commence with the highest quality format, and then see if the same audio is retrieved when mp3 encoded, or at low sample rate, or using any format which introduces error. If one starts with the highest quality audio, one can then find the point at which retrieval is affected. Robustness cannot properly be determined if one does not have access to the initial recordings.

2.4.3. Rich Data

Master recordings usually consist of many separate tracks. Each track is a high quality recording, usually of a single instrument recorded during a single session. Access to such material would provide researchers with very rich data to analyse, and retrieval experts with many more methods by which to perform retrieval.

Many analysis and processing routines performed on audio data are, in effect, reverse-engineering of master recordings. Source separation attempts to separate the individual voices. Instrument recognition attempts to identify the various instruments used in a recording[6]. Onset detection and note recognition techniques are plagued by complexities due to the polyphonic, multi-voice nature of most recordings[7]. Yet on most masters, the instruments are separated on different tracks, the voices are on different tracks, and each track has few polyphonies. Transcription becomes easier since it need only be applied to one voice under known conditions.

Furthermore, effects are introduced in the mixing process. Fading, time stretching and distortion may be added in the mix but the original master tracks remain untouched. Using the master tracks in the testbed would allow one to retrieve audio based on a more meaningful musical similarity measure, since most of the audio production-based (dis)similarity would not be present.

2.4.4. Metadata

Master recordings also contain a significant amount of metadata. For studio recordings produced on DAWs, which includes the vast majority of commercial recordings, each track is labelled with meaningful metadata, such as a timestamp, title, performer, instrument, and so on. The exchange formats mentioned in Section 1.4, such as AAF and AES31, also include editing and mixing information so that the final production mix can be recreated from the master tracks.

This, together with the innate richness of the data, provides powerful tools which can be exploited by MIR systems. One can, for instance, search for all of a performer's guitar tracks, or for a specific percussive style that may occur on any recording, or all uses of a popular sample. Not to use master recordings would be to throw away, meaningful metadata for which it would be impossible to recreate.

3. USABILITY GUIDELINES

For a testbed to be usable by the entire MIR, it should not require the MIR community to adapt to its preferred format, platform, operating system, or development environment. Instead it should support all popular variations and provide a mechanism whereby users of unusual variants may still access the testbed. This is for the simple reason that researchers will not use a system which requires them to adapt to a new and possibly unproductive environment. Furthermore, requiring many researchers to each support one specific option is

duplication of effort. It is more efficient to build support for many alternatives directly into the testbed.

In order to achieve this, open standards are required wherever possible. Proprietary formats may not be supported by common audio players and streamers, or many development environments, e.g., GNU C++ libraries do not provide WMA decoders. These open standards are also necessary because the testbed is there to help the community. If a proprietary standard is used and an MIR system fails, it may not be possible to tell why it failed. Without understanding an encoding technique, one cannot determine why that encoder may have caused an audio file not to be retrieved.

Finally, closed proprietary formats cause *lock-in*. That is, users of the testbed will then need to use the encoders and decoders associated with this format, as well as the associated processing tools. These are often all originating from the same company. Productivity then necessitates a reliance on this company, and other alternatives are rejected because they cannot be used with the existing audio files. Rather than building a testbed for the wider MIR community, this will create a software specific testbed of only limited, specialized use.

4. COMPLEXITY GUIDELINES

The MIR community is incredibly diverse. It is comprised of musicologists, signal processing engineers and library scientists, to name just a few. Thus, users of a testbed are often highly specialized. They may have limited knowledge of music theory, programming, information retrieval or digital signal processing. A format reliant on an arcane metadata format will be impractical for engineers, and similarly, audio formats using sophisticated psychoacoustic models will seem obtuse to information retrieval experts.

The solution is to implement simple, transparent and well-understood formats wherever possible. Metadata should be encoded in text format so that it can be simply read without requiring advanced programming skills. Audio encoding should not require advanced processing or prior knowledge of acoustics or human sound perception. The format should be one that is either supported by most languages and development environments, or one where it is easy to construct decoders and encoders. Wherever possible, converters should be provided so that the audio file can be played or analysed in all major formats. Such a simple scheme allows for the entire MIR community to benefit from the testbed with minimal time spent on acquiring irrelevant format-specific skills.

5. RECOMMENDATIONS

In this document we have outlined the options available for the audio files used in an MIR testbed and music digital library. We have also provided and justified a set of guidelines for the audio files, formats and wrappers used in the testbed construction. Given these options and guidelines, it is now possible to list a set of explicit recommendations concerning formats, converters, quality and copyright infringement prevention. These recommendations come with the caveat that the primary factor in determining the nature of audio files in a testbed is restricted by what the copyright holders are willing to provide. Nevertheless, the creators of the testbed and the MIR/MDL infrastructure experts should be able to effectively argue the case for the preferred audio files under the preferred conditions.

Recommendation 1: A popular, simple, well-understood and uncompressed format should be used as the primary format for encoding audio files.

We noted earlier that WAV and AIFF are the two most popular raw audio formats and are supported by almost all DAWs and audio players. However, almost all development environments offer encoding and decoding of WAV files, whereas AIFF support is not built-in to some development environments (LabWindows and Labview, for instance). In addition, standards bodies such as the EBU and the AES have formally endorsed WAV. For these reasons, we recommend that uncompressed WAV files be the main audio format.

Recommendation 2: Whenever possible, the master recordings should be obtained and stored with any metadata and audio production information.

The reasons for use of master recordings were outlined in Section 2.4. They represent the highest quality audio available, and the closest approximation to the groundtruth. Almost all digital masters are stored as, or can be easily converted to Broadcast WAV format. The AES31 standard, which is supported by DAW manufacturers and endorsed by the audio production and broadcasting communities, provides an open and simple standard for easily transferable Broadcast WAV encoded audio files and associated metadata. Together with Recommendation 1, this provides a ringing endorsement for the use of Broadcast WAV format, raw audio master recordings as the testbed *essence*.

Recommendation 3: Testbed creators must guarantee that files can be analysed using all popular development environments, listened to with all popular audio players, and on all major operating systems.

This necessitates system testing by MDL designers, but should not require effort on the part of MIR system researchers. Popular development environments include the analysis software MATLAB (popular in the signal processing community), the programming languages Java and C/C++, and the scripting language Perl (useful for informal programming). Relevant audio players include Quicktime, Windows Media Player, RealPlayer and WinAmp. Since some popular audio formats are not supported by all major media players, it may be necessary to provide converters. Again, this should be implemented on the testbed side, not by the individual MIR researchers. The operating systems that should be supported are Mac, Windows and Linux/UNIX. Support should extend to recent versions, not just the current version, e.g., Mac OS 8.x and 9.x as well as OS10.

Recommendation 4: The testbed should allow multiple formats. Although the first three recommendations suggest a preferred audio format and its support, they do not preclude the use of multiple formats in the testbed. Multiple formats should be used for the storage of audio because it allows one to skip the audio conversion step where it would be used, because audio files may be provided in different formats, and because it provides researchers with a rich and heterogeneous testbed that allows evaluation of diverse retrieval systems.

Recommendation 5: MIR researchers must be allowed to listen to the material in the testbed. Any artifacts or distortions introduced to satisfy the demands of copyright holders should not restrict the ability of researchers to analyse their MIR system and evaluate its performance on the corpus.

This recommendation depends on the restrictions imposed by copyright holders and the reaction of the MIR community to those restrictions. In order to ensure that no high quality audio is leaked outside the research community, severe limitations will most likely be placed on the ability to listen to the files in the testbed. Nevertheless listening tests are an essential part of music-related research.

At a minimum, researchers should be allowed to listen to a low quality popular format version of the audio with embedded artifacts. Options include streaming, providing audio in mono, in a highly compressed form, embedding artifacts such as pings and drop-outs, thumbnailing and watermarking. Streaming seems reasonable although it is possible to *rip* an audio stream and redistribute it as a file. Listening to artifacts can be irritating and detract from the ability to perform proper evaluation in a listening test, as do all audio modifications. Furthermore, they would prove unsuitable for any demonstrations of an MIR system which utilises the testbed. Thumbnailing is still at the forefront of research, and a thumbnail may not contain the most relevant audio material.

Watermarking is also problematic because it entails emphasised responsibility and enhanced liability. Any leaked audio can be tracked back to audio researchers, or atleast to the testbed user community. This implies that the maintainers of the testbed can more easily be held liable since it can be shown that leaked material originated from the testbed. Due to these issues, watermarking is discouraged.

It is difficult to gage in advance how much of an imposition any of these limitations will place on research. However, all artifacts and distortions can affect the evaluation of relevance. Furthermore, these artifacts are innately problematic in listening tests since they affect the way audio sounds. Therefore we recommend that they be avoided wherever possible.

6. CONCLUSION

The guidelines and recommendations presented here are intended for the creation of the most powerful and accessible music digital library possible. Throughout, we have assumed that size constraints on the testbed are minimal. This allows us to recommend the high quality uncompressed master recordings in favour of small, highly compressed files such as mp3s. If there are severe limitations on testbed size, then all of the above recommendations would need revision. The size of other data in the testbed, such as symbolic music representations, would need to be taken into account. However, it is a reasonable assumption that any large-scale testbed intended for use by the greater MIR community would have ample space for all data.

Issues pertaining to format, quality and copyright infringement prevention are not unique to MIR and music digital library creation. The image and video retrieval communities have dealt with similar issues for years. Quality is not as strong an issue for both media, since almost all video is high quality, and high quality images are easily found. The image community primarily uses easily interchangeable uncompressed image formats, and the video community has yet to settle on any standard, although MPEG-2 is common. But both communities are plagued by the same copyright issues which affect MIR. The multimedia retrieval research communities have various projects underway to provide large testbeds of material with few copyright access issues.

Finally, the question of preferred audio format has been tackled in the related discipline of audio restoration and preservation. Almost universally, the members of this community recommend the storage of files in Broadcast WAV format (AES, EBU, Audio Restoration Services, the Library of Congress's National Digital Library Program, …). Thus, adoption of such a format will allow the MIR research community to easily collaborate with this and other related disciplines.

7. REFERENCES

[1] "Specification of the Broadcast Wave Format, EBU Tech. Doc. 3285," European Broadcast Union, Geneva, Switzerland July 1997

[2] "AES standard for network and file transfer of audio," Audio Engineering Society 1999,2001. www.aes.org/standards/b_pub/aes-standards-in-print.cfm

[3] *Metadata applications and management*, International Yearbook of Library and Information Management, 2003-2004. London: Facet Publishing, 2004.

[4] T. Thiede, *et al.*, "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, pp. 3-29, 2000.

[5] M. Erne, "Pre-Echo Distortion," in *Audio Coding Artifacts Educational CD-Rom*, J. Herre and J. D. Johnston, Eds.: Audio Engineering Society, 2001.

[6] P. Herrera, *et al.*, "Towards Instrument Segmentation for Music Content Description: A critical review of instrument classification techniques," presented at ISMIR, Plymouth, MA, 2000.

[7] J. Pickens, *et al.*, "Polyphonic Score Retrieval," presented at the 3rd ISMIR, Paris, France, 2002.