A Framework for the Automatic Description of Musical Structure Using MPEG-7 Audio

Elaine Smyth, Kevin Curran, Paul Mc Kevitt, and Tom Lunney

School of Computing and Intelligent Systems
Faculty of Engineering
University of Ulster, Magee
BT48 7JL, Derry/Londonderry, Northern Ireland
E-mail: {smyth-e1, kj.curran, p.mckevitt, tf.lunney}@ulster.ac.uk

Abstract. A great deal of research has been conducted in the area of musical analysis. However, very little has been done with respect to utilising the MPEG-7 metadata standard to describe the structural content of music at a fine level of granularity. The aim of this paper is to provide an overview of the field of structural musical analysis and to present an architecture for the AMUSED system for automated structural analysis and description using the MPEG-7 standard.

1. Introduction

A great deal of research has been conducted in the area of musical analysis, both non-automated and automated, and some with particular emphasis on musical structure. However, very little has been done with respect to using the relatively new MPEG-7 standard as a basis for the description of hierarchical musical structure. MPEG-7 is particularly useful as it produces a standardised output which can potentially be used within other systems. We are developing a system called AMUSED (Automated MUsical StructurE Description) which performs real-time structural analysis using the MPEG-7 standard as a basis for the identification and the subsequent description of musical structure. Output from AMUSED will be in the form of an MPEG-7 compliant XML file. This output will be useful to a number of applications which seek to recognise musical structure, such as interactive music performance, visual based music applications or error concealment modules within streaming media algorithms, among other applications. This paper aims to provide an overview of previous research in the field of musical analysis and the use of XML in music notation, leading up to an introduction to the MPEG-7 standard and presentation of the proposed Automated MUsical StructurE Description (AMUSED) system architecture for identification and description of musical structure using MPEG-7 Audio.

2. Automated Music Structure Analysis

The automated discovery of patterns within music is an important problem within computational music analysis. Patterns emerge from repetitions within the music itself, and these patterns and repetitions are generally a good indication to musical structure. Repetitions can range in size from simple repeating note sequences, to recurring sections or phrases within a musical piece. In much the same way, structure

can be defined at differing levels of granularity based on these patterns and repetitions; from high level song form down to the single note level, if required. In addition, each of these structures in turn can be interrelated and higher level structures can contain complex organization within themselves.

There have been many techniques developed to discover patterns and structure within music which have resulted in a number of varying and sometimes complex approaches. To outline them all here would be a lengthy process, therefore only a brief overview of a selection of these approaches will be provided. Patterns have been extracted in variety of ways; using multi-dimensional datasets [19], a String-Joining approach [13], matrices [1], multiple viewpoints and a suffix tree data structure [6]. Some even allow the discovery of patterns in retrograde and/or inversion [21], and some simply extracts patterns from polyphonic MIDI sequences [20]. Perhaps the most novel approach is that in [5]; they use what they call a Spiral Array model for recognising and visualising tonal patterns based on pitch/time structures. Of the approaches mentioned thus far none actually deal with the identification of structure, only pattern recognition; however, this is a crucial element within the eventual identification of structure. The approaches mentioned next not only identify patterns, but also infer some level of coarse musical structure.

Foote and Cooper [10] use a somewhat original matrix style approach to music structure visualisation. They represent a piece of music as a 2-dimensional matrix. Within this matrix levels of self-similarity at any one point are indicated by light (similar) and dark (dissimilar) shading. High level structure can be inferred from the matrix in the form of intro, verse, chorus, etc. Dannenberg and Hu [7], Lu and Zhang [18], and others also utilise a matrix approach which result in the high level identification of structure. Finally, Vercoe and Chai [27], present a system that can automatically analyse the repetitive structure of musical signals; structure is discovered through repeating segments within a piece of music. The output from their system is, again, a relatively high level representation of structure in the form of ABBA, for example.

Most of these systems appear to work sufficiently well in identifying patterns and, where applicable, structure. However, they all suffer the inability to produce some level of universally standardised output, somewhat limiting their use outside their current project scope. In addition, those which do aim to identify structure do so at a relatively high, non-hierarchical level. Furthermore, the actual representation of music varied from system to system; e.g. strings, vectors, matrices and even spheres. They were also based on a variety of musical information; some on pitch, duration, interval, and others on spectral energy, rhythm and timbre. In addition, some required pre-processing before analysis could be performed, which is not ideal for use with real-time processing environments. It would be extremely valuable to have a system which analyses and represents all musical information in a standardised way; the output of which would be useful to any number of applications, even those outside

music theory. Wiggins, et al [28], go some way towards providing a standard representation, but it is not on par with the complex representation possible with XML-based notation. In addition, XML is already standardised and can be easily parsed for specific information because it is text-based.

3. XML and Music Notation

XML (eXtensible Markup Language) is a standardised, text-based method for describing the structure of data. It provides a very rich schema for defining complex documents and data structures. Text-based representations are ideal for music. Music does not lend itself to easy interpretation by computer, unlike language which can be easily represented and understood by computers because of the ANSI standard. Of particular benefit to music applications is the fact that XML allows support for multiple descriptions of the same data, and hierarchies are as fundamental to XML as they are to music notation.

There are a variety of XML-based markup projects in existence that are specifically aimed at music. XScore [11], Music Encoding Initiative (MEI) [22], MusicXML [26], and Music Content Markup Language (MCML) [25], to name a few. Although all are based on XML, they do have varying purposes. E.g. XScore is an application of XML for describing the musical score, and MCML was developed as part of another project (MIDILib) for content-based queries and navigation. While MEI strives to meet a broader range of music applications, it also avoids the confusion of other XML standards by using familiar names for elements and attributes, e.g. <note> and <chord>. MusicXML is perhaps the most mature endeavour for music encoding using XML. It has its roots in academia and has made its way into a number of commercial applications and has even been put forward for standardisation.

There are a few tools which harness the power of XML-based markup languages; as mentioned, the MIDILib project utilizes MCML; in addition, MiXA, VExPat and Sharpeye, among others, make use of MusicXML. MiXA is a web based musical annotation system [17], whereas VExPat is pattern extraction system which uses the MusicXML representation of music as a basis for analysis [24]. Sharpeye is a music reader which converts a scanned image of printed music into a MIDI, NIFF or MusicXML file [16]. There do not appear to be any XML applications which deal explicitly with the automatic identification and representation of structure within music.

A potential drawback is that most current DTDs restrict themselves to Common Western Music Notation (CWMN), with some including tablature. MusicXML, e.g. was designed to represent musical scores and sheet music, specifically common western musical notation from the 17th century onwards [26]. MPEG-7 is a further XML-based standard, but is more general in its approach; it is not based on the representation of sheet music or scores, but can deal with the representation of music directly from the audio file; in addition it can also deal with spoken content. Analysis of the sound file directly frees the representation somewhat from the restrictions of a

specific notational approach, like CWMN. Furthermore, it allows for significantly more detailed information to be stored about a particular piece of music.

4. Overview of MPEG-7 Audio

The MPEG-7 standard currently consists of 8 parts, part 4 of which deals specifically with the description of audio data; formally recognised as standard 15938-4. The main MPEG-7 elements are Descriptors (D), Description Schemes (DS), and a Description Definition Language (DDL). Ds are intended to describe low-level audio features; they are the building blocks of the system. DSs are designed to describe higher level audiovisual features. DSs produce more complex descriptions by combining multiple Ds and DSs and declaring relationships among the description components. The DDL provides the descriptive foundations through which users can create their own Ds and DSs [23].

The Audio Framework tools are applicable to the description of general audio; a graphical representation of the Framework is provided in Fig. 1. The generic Audio Framework contains low-level descriptors designed to provide a basis for the construction of higher level audio description schemes. The Low-Level Descriptors (LLDs) permit the description of an audio signal's spectral and temporal features.

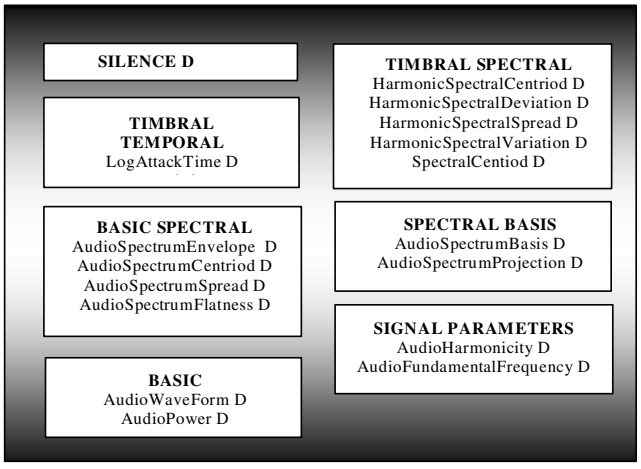

Fig. 1. Audio Framework [14]

In the first version of the standard there were seventeen LLDs for general use in a variety of applications. Since 2004, there has been an extension to the original standard (Amendment 1) to include additional Ds for such things as background noise level, balance, bandwidth, etc, with proposals for a second amendment to include Ds

to describe audio intensity, rhythmic patterns, chord patterns, and so forth [15], [12]. In addition to the LLDs there are five general sets of high-level audio description tools, which aim to encompass some application areas, such as sound recognition, musical instrument timbre, spoken content, melodic contour and melody. These specialised tools may be used in conjunction with the other tools within the standard. The high-level tools provide both functionality and also serve as examples of how to use the low-level framework [14].

MPEG-7 Ds and DSs have been successfully utilised by a number of projects over the last few years in a variety of application areas [2], [4], [8], [9]. These can be broadly categorised into one or more of the following areas: Query-By-Example (QBE), fingerprinting for audio identification, indexing and archiving, MPEG-7 authoring tools, audio analysis, audio classification, and Digital Rights Management (DRM). Musicstructure.com is perhaps the only existing MPEG-7 based application which deals solely with the analysis and representation of musical structure, but again this system only performs partitioning into a relatively high level structure [3]. MPEG-7 holds great promise in relation to the useful description of audio. It will have many applications, both within music theory and within information search and retrieval, as well as many as yet unknown application areas.

5. AMUSED System Architecture

The Automated MUsical StructurE Description (AMUSED) project seeks to overcome the inadequacies identified and provide a novel approach for the automated extraction, analysis and description of digital audio structural characteristics. It aims to develop a Java-based real-time architecture which will take digital musical audio (in most common formats) as input and produce standardised MPEG-7 compliant descriptions of musical structure as output. A diagram representing an overview of the proposed system architecture is shown in Fig. 2. The main focus will be on the accurate description of musical structure, based on within-song similarity, at a fine enough granularity to make it universally useful to all manner of applications; e.g. use within a visual based music application or an error concealment module within a streaming media application, or for use within music theory or musical analysis. Since the MPEG-7 output file is in a standardised format (XML) it creates the possibility of any system so designed being able to hook into and use it for further processing.

A digital audio signal is received into the system as a data stream. Spectral analysis is performed on the incoming data stream to accumulate low level information about the audio signal; signal power, spectrum spread, fundamental frequency, etc. This information is ‘extracted’ from the signal and described using a selection of the MPEG-7 Audio LLDs. Extracting all the Low-Level Descriptors (LLDs) would be a wasteful use of processing resources, therefore, a careful selection will be made with the project objectives in mind. Only those Descriptors which will be useful in pattern discovery, comparison and structural analysis will be chosen, as well as those with the ability to, thereafter, accurately describe structure. For example, the *SoundModelStateHistogram* is used to compare sound segments using histograms of

their state activation patterns, thus it will be useful for self-similarity comparisons, segmentation and structural analysis. The *SoundModelStatePath* could be a good candidate for modelling audio data at a micro level. The series of states could be used as 'markers' within an audio stream to identify particular elements of interest. It is proposed that perhaps a speciality Description Scheme be constructed from a choice of Descriptors in part 5 of the MPEG-7 standard. Spectral representations have the potential for accurately identifying lower level structure due to their relationship with the actual power of the signal. The power of the signal is the lowest level representation of audio therefore any descriptor which corresponds to this will demonstrate some form of low level structure over time.

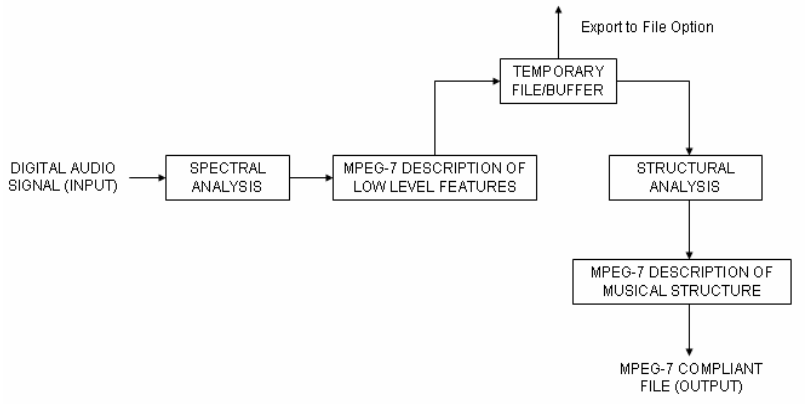

Fig. 2. Overview of Proposed System Architecture

Obviously some type of buffer/temporary file will be required to store the accumulative information produced by the Low Level Description module prior to their analysis by the structural analysis module. This file will be held in working memory and will also have the option of being exported as an XML file at this stage. Following on from the Low Level Description module, the structural analysis module performs further analysis to identify patterns, repetition and regions of self-similarity within the information contained within the extracted LLDs. This information will then used as a basis for the construction of a hierarchical model of musical structure, which is consequently described by the MPEG-7 structural description module. The final output from the system is in the form of an MPEG-7 compliant XML file, which can potentially be used within other systems.

6. Conclusion

This paper provided a review of previous research and an overview of the MPEG-7 standard, together with a brief synopsis of the proposed system architecture for the

AMUSED system. Many of the systems reviewed lack a standardised output and are therefore not particularly useful outside their own project scope. The use of XML for music notation is a step in the right direction; however, the projects outlined did not capture anything that significant about the music. E.g. XScore describes the written musical score which is really only useful for archiving and data exchange. Built on top of the XML base is the MPEG-7 standard. MPEG-7 offers a standardised scheme for the description of audio, among other media. Some applications are already in existence which utilise MPEG-7; however, very few offer the automated description of musical audio structure at a fine level of granularity. Indeed, most are QBE-style applications. The AMUSED system aims to overcome the inadequacies identified and provide a novel system for the automated extraction, analysis and description of digital audio structure and produce standardised, MPEG-7 output.

References

1. Aucouturier, J., and Sandler, M.. *Finding Repeating Patterns in Acoustic Musical Signals*, AES, 22nd International Conference on Virtual Synthetic and Entertainment Audio, Espoo, Finland, (2002), 145-152
2. Batke, J., Eisenberg, G., Weishaupt, P., Sikora, T. *A Query By Humming System using MPEG-7 Descriptors*, AES Convention Paper 6137, AES 116th Convention, Berlin, Germany, (2004)
3. Casey, M. Musicstructure.com, http://www.musicstructure.com/intro.html
4. Celma, O., Gomez, E., Janer, J., Gouyon, F., Herrera, P., Garcia, D. *Tools for content-based retrieval of audio using MPEG-7: the SPOffline and the MD Tools*, AES 25th International Conference, London, (2004)
5. Chew, E. *MuSA: Music Information Processing*, Proceedings of the 2nd International Conference on Music and Artificial Intelligence, Edinburgh, Scotland, (2002), 18-31
6. Conklin, D. and Anagnostopoulou, C. *Representation and Discovery of Multiple Viewpoint Patterns*, International Journal of New Music Research, Vol. 24, No. 1, (2001), 51–73
7. Dannenberg, R. and Hu, N. *Pattern Discovery Techniques for Music Audio*. ISMIR 2002 - 3rd International Conference on Music Information Retrieval, IRCAM – Centre Pompidou Paris, France, (2002)
8. Dumouchel, P., *MPEG-7 Audiovisual Document Indexing System (MADIS*), Testbed Development, RISQ 2003, 14th Edition, Annual Event of the Network of Scientific Information of Quebec (RISQ), Annual Workshop of CANARIE, Montreal, Canada, (2003)
9. Eisenberg, G., Batke, J., Sikora, T. *BeatBank – An MPEG-7 Compliant Query by Tapping System*, AES Convention Paper 6136, AES 116th Convention, Berlin, Germany, (2004)
10. Foote, J. and Cooper, M. *Visualizing Musical Structure and Rhythm via Self-Similarity*. International Conference Computer Music, Habana, Cuba, (2001), 140-148
11. Grigaitis, R. eXtensible Score Language (XScore) 0.01. http://grigaitis.net/xscore
12. Gruhne, M. *MPEG-7 Audio.* Workshop Music Network, Barcelona, Spain, (2004)
13. Hsu, J., and Liu, C. *Discovering Nontrivial Repeating Patterns in Music Data.* IEEE Transactions on Multimedia, Vol. 3, No. 3, (2001), 43-52
14. ISO/IEC. *Information Technology – Multimedia Content Description Interface – Part 4: Audio.* ISO-IEC JTC 1/SC 29/WG 11, ISO-IEC FDIS 15938-4:2001(E), http://projekt.rz.tu-ilemnau.de/~kfn/seminar13/MPEG-7-Dokument.pdf, (2001)
15. ISO/IEC. *Information Technology – Multimedia Content Description Interface – Part4: Audio*, Amendment1:Audio Extensions, ISO/IEC JTC1/SC29/WG11 N4769, http://www.itscj.ipsj.or.jp/sc29/open/29view/29n4854t.doc, (2002)
16. Jones, G., (2005), *SharpEye Music Reader*, http://www.visiv.co.uk/about.htm

17. Kaji, K., and Nagao, K., *MiXA: A Musical Annotation System*, Proceedings of 3rd International Semantic Web Conference, Hiroshima, Japan, (2004)
18. Lu, L. Wang, M., and Zhang, H. *Repeating Pattern Discovery and Structure Analysis from Acoustic Music Data*, Proc. of IEEE International Conference on Multimedia and Expo (ICME '04), Taipei, Taiwain, (2004)
19. Meredith, D., Lemstrom, K., and Wiggins, G. *Algorithms for discovering repeated patterns in multi-dimensional representations of polyphonic music*. Journal of New Music Research, Vol. 28, No. 4, (2003), 334–350
20. Meudic, B. *Automatic pattern extraction from polyphonic MIDI files, Les* Journees d'Informatique Musicale, 10th Edition, L'ecole Nationale De Musique Du Pays De Montbeliard, (2003)
21. Ren, X., Smith, L., and Medina, R. *Discovery of Retrograde and Inverted Themes for Indexing Musical Scores*, ACM/IEEE Joint Conference on (JCDL'04), Tucson, AZ, USA**,** (2004), 252-253
22. Roland, P. *The Music Encoding Initiative (MEI)*, Proceedings of 1st International Conference Musical Applications using XML, Milan: State University of Milan, (2002), 55–59
23. Salembier, P., and Smith, J. *MPEG-7 Multimedia Description Schemes*, IEEE Transactions Circuits and Systems for Video Technology, Vol. 11, No. 6, (2001), 65-73
24. Satana, H., Dahia, E. L., and Ramalho, G. *VExPat: An Analysis Tool for the Discovery of Musical Patterns*, Proceedings of IX Brazilian Symposium on Computer Music, Campinas, SP, (2003)
25. Schimmelpfenning, J and Kurth, F. *MCML: Music Contents Markup Language*, Proceedings of International Symposium of Music Information Retrieval, Plymouth, MA, USA, (2000)
26. Stewart, D. *XML for Music*, http://emusician.com/mag/desktop/emusic_xml-music/
27. Vercoe, B., and Chai, W. *Structural Analysis of Musical Signals for Indexing and Thumbnailing*, ACM/IEEE Joint Conference on Digital Libraries, New York, (2003), 27– 34
28. Wiggins, G., Harris, M., and Smaill, A. *Representing Music for Analysis and Composition*, Proceedings of the 2nd IJCAI AI/Music Workshop, (1989)