WEMIS 2009

Workshop on

Exploring Musical Information Spaces

In conjunction with ECDL 2009

Corfu, Greece, October 2009

ISBN: 978 - 84 - 692 - 6082 - 1

Foreword

There is an increasing interest towards music stored in digital format, which is witnessed by the widespread diffusion of standards for audio like MP3 and of web-based services to listen or purchase music. There is a number of reasons to explain such a diffusion of digital music. First of all, music crosses the barriers of national languages and cultural backgrounds and can be shared by people with different culture. Moreover, music is an art form that can be both cultivated and popular, and sometimes it is impossible to draw a line between the two, for instance in the case of jazz or of ethnic music. These reasons, among others, may explain the increasing number of projects involving the creation of music digital libraries. A music Digital Library (DL) allows for, and benefits from, the access by users from all over the world, it helps the preservation of cultural heritage, and it is not tailored only to scholars' or researchers' needs.

The availability of music collections to a wide number of users, needs to be paired by the development of novel methodologies for accessing, retrieving, organizing, browsing, and recommending music. The research area devoted to this aspect is usually called Music Information Retrieval (MIR) although retrieval is only one of the relevant aspects. Given the particular nature of music language, which does not aim at describing objects or concepts, typical metadata give only a partial description of music documents. Thus great part of MIR research is devoted to content-based approaches, aimed at extracting relevant descriptors, computing perceptually based similarity measures, identifying music genres and artists, naming unknown songs, and recommending relevant music items.

In recent years, most of the published research results focused on the extraction of relevant features from audio recordings, aimed at providing tools for retrieval of and access to popular music for music consumers. Yet, we believe that there is still the need of a forum devoted to the investigations of new paradigms of interacting with music collections for a wider variety of users, including musicologists, theorists, and music professionals, in order to promote the dissemination of cultural heritage. This goal is achieved by means of research on the formalization of users' needs, on novel paradigms for browsing personal collections and music digital libraries, and on the role that music theory plays on the concepts of relevance and similarity. To

this end, music content should thus include the symbolic representation of music works and the information on the structure of music pieces.

This document contains the 13 papers presented at WEMIS 2009, held in Corfu, Greece, October 1-2 , 2009. We would like to express our sincere gratitude to the ECDL' 09 organizers, and specially to the members of the program committee for ensuring a high quality proceedings.

This document can be found at:
`http://www.dlsi.ua.es/wemis09/proceedingswemis2009.pdf`

Nicola Orio

Andreas Rauber

David Rizo

Workshop Chairs

Nicola Orio, University of Padova, Italy

Andreas Rauber, Vienna University of Technology, Austria

David Rizo, University of Alicante, Spain

Program Committee

George Tzanetakis (University of Victoria, Canada)

Kjell Lemström (University of Helsinki, Finland)

Darrell Conklin (City University, London, UK)

José Manuel Iñesta (University of Alicante, Spain)

Carlos Agón (IRCAM, Paris, France)

David Bainbridge (University of Waikato, New Zealand)

Roberto Bresin (Stockholm University, Sweeden)

Sergio Canazza (Università degli Studi di Udine, Italy)

Giovanni De Poli (University of Padova, Italy)

Gert Lanckriet (University of California, San Diego, USA)

Olivier Lartillot (University of Jyväskylä, Finland)

Massimo Melucci (University of Padova, Italy)

Jeremy Pickens (FX, Palo Alto, USA)

Simon Dixon (Queen Mary University of London, UK)

Anssi Klapuri (Tampere University of Technology)

Francois Pachet (Sony CSL Paris, France)

Emilios Cambouroupoulos (Aristotle University of Thessaloniki, Greece)

Stephen Downie (University of Illinois at Urbana-Champaign, USA)

Organized By

Department of Information Engineering, University of Padova

MIR Group, Vienna University of Technology

GRFIA. Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante

Aknowledgements

The organization of the workshop has been partially supported by a grant of the University of Padova for the project "Analysis, design, and development of novel methodologies for the study and the dissemination of music works", and by the Spanish Ministry projects: TIN2006--14932--C02 and Consolider Ingenio 2010 (MIPRCV, CSD2007-00018), both partially supported by EU ERDF.

TABLE OF CONTENTS

The Vicentini sound archive of the Arena di Verona..... 1

Federica Bressan, Sergio Canazza, Daniele Salvati

Timbre Similarity Search with Metric Data Structures..... 7

Francisco Costa, Fernanda Barbosa

A location-tracking interface for ethnomusicological collections..... 12

Michela Magas, Polina Proutskova

Towards the Disintermediation of Creative Music Search: Analysing Queries To Determine Imporant Facets..... 18

Charlie Inskip, Andy MacFarlane, Pauline Rafferty

Preserving today for tomorrow: a case study of an archive of Interactive Music Installations..... 24

Federica Bressan, Sergio Canazza, Antonio Rodà, Nicola Orio

Multi-modal Analysis of Music: A large-scale Evaluation..... 30

Rudolf Mayer, Robert Neumayer

metamidi: a tool for automatic metadata extraction from MIDI files..... 36

Tomás Pérez-García, José M. Iñesta, David Rizo

Integration of Chroma and Rhythm Histogram Features in a Music Identification System..... 41

Riccardo Miotto, Nicola Montecchio

Ensemble of state-of-the-art methods for polyphonic music comparison..... 46

David Rizo, José M. Iñesta, Kjell Lemström

Measuring Harmonic Similarity Using PPM-based Compression Distance..... 52

Teppo E. Ahonen

The NEUMA Project: Towards Cooperative On-line Music Score Libraries..... 56

L.Abrouk, H. Audéon, N. Cullot, C. Davy-Rigaux, Z. Faget, D. Gross-Amblard, H. Lee, P. Rigaux, A. Tacaille, E. Gavignet, V. Thion-Goasdoué

Analytic Comparison of Audio Feature Sets using Self-Organising Maps..... 62

Rudolf Mayer, Jakob Frank, Andreas Rauber

Matching Places of Interest With Music..... 68

Marius Kaminskas, Francesco Ricci

The Vicentini sound archive of the *Arena di Verona* Foundation: A preservation and restoration project

Federica Bressan
Department of Computer Science
University of Verona
Strada le Grazie 15, 37134 Verona
Email: federica.bressan_01@univr.it

Sergio Canazza
Lab. AVIRES
University of Udine
Via Margreth 3, 33100 Udine, Italy
Email: sergio.canazza@uniud.it

Daniele Salvati
Lab. AVIRES
University of Udine
Via Margreth 3, 33100 Udine, Italy
Email: kabit@tiscali.it

***Abstract*—In the sound archive field, a long-term maintenance of the collective human memory in its original form is not sustainable. All physical carriers are subject to degradation and the information stored on such carriers is bound to vanish. Only a re-mediation action can prevent precious knowledge from being permanently lost. We present the first results carried out by a joint cooperation project between the University of Verona and the Arena di Verona Foundation, with the scientific support of Eye-Tech (Udine, Italy), which aims at establishing a HW/SW platform with the purpose of preserving and restoring the Vicentini archive of the Arena Foundation.**

I. Introduction

The opening up of archives and libraries to a large telecoms community represents a fundamental impulse for cultural and didactic development. Guaranteeing an easy and ample dissemination of some of the fundamental moments of the musical culture of our times is an act of democracy which cannot be renounced and which must be assured to future generations, even through the creation of new instruments for the acquisition, preservation and transmission of information. This is a crucial point, which is nowadays the core of the reflection of the international archive community. If, on the one hand, scholars and the general public have begun paying greater attention to the recordings of artistic events, on the other hand, the systematic preservation and access to these documents is complicated by their diversified nature and amount.

Since the paper used in 1860 (first audio recording by Édouard-Léon Scott de Martinville "Au Clair de la Lune" using his phonautograph) to the modern Blu-ray Disc, what we have in the audio carriers field today is a Tower of Babel: a bunch of incompatible analog and digital approaches (paper, wire, wax cylinder, shellac disc, film, magnetic tape, vinyl record, magnetic and optical discs, etc.) without standard playback systems able to read all of them. Be it in the analogue or digital domain, audio memories are stored on different types of carriers that are equally subject to a process of physical decay. In order to prevent the information from being eroded or completely erased, it is necessary to keep transferring the signal to new carriers.

As far as audio memories are concerned, preservation is divided into passive preservation, meant to defend the carrier from external agents without altering the structure, and active preservation, which involves data transfer on new media. Active preservation allows us to copy, handle and store the audio documents in virtual spaces that can be remotely accessed by large communities. Digitization is necessary to prevent the documents from disappearing, and it is desirable because it allows to distribute them on a wide scale. The commingling of a technical and scientific formation with historic-philological knowledge also becomes essential for preservative re-recording operations, which do not completely coincide with pure A/D transfer, as is, unfortunately, often thought. Along the years some guidelines have been sketched [1], [2]. The REVIVAL project is concerned with both passive and active preservation, although this paper focuses on the latter.

With its tens of thousands of audio documents, some of which extremely rare or even unique, the Vicentini archive of the Arena Foundation (Verona, Italy) has potential to gain highest cultural interest worldwide. In spite of this potential, a reconnaissance mission evidenced that a large part of the archive is at risk of vanishing, with a serious damage for the musical heritage existing in Europe: for this reason, a joint cooperation project (REstoration of the Vicentini archive In Verona and its accessibility as an Audio e-Library: REVIVAL) between the University of Verona and the

Arena Foundation, with the scientific support of Eye-Tech (Udine, Italy) started.

Sect. II describes the REVIVAL project objectives and the audio preservation/restoration laboratory created in Arena. After an overview of the Vicentini's archive in Arena Foundation (Sect. III), the first results of the project are presented: the audio preservation protocol (Sect. IV), the different approaches that can be adopted in the access copy (Sect. V) and an innovative audio restoration algorithm developed on purpose, i.e. for the particular characteristics of the Vicentini audio documents (Sect. V-A).

II. Revival project

The devising of modern audio restoration operational protocols aims at avoiding the superposition of modernized phonic aspects during the re-recording procedures, which may distort the original audio content. This requirement is essential for a philologically correct preservation, that needs to be supported with all the knowledge that can be acquired with and aside the signal. This knowledge lies in the history of the compositional thought and of the musical technology connected with it. Audio preservation should be ascribed in the more general scientific field of musical analysis only when it is consistent with such principles.

Concerning REVIVAL, the first task was devoted to the development of an operational protocol specifically finalized to the preservation of the audio documents held by the Vicentini archive (described in Sect. IV). At the same time, a preservation/restoration laboratory was established at Arena Foundation, enabled by tape dryers and custom, non-invasive disc and tape players. An example of non-invasice player is the Photos of GHOSTS system [3], developed by the authors, able to extract audio data from recorded grooves of a shellac disc, acquired using an electronic camera. The images can be processed to extract the audio data also in the case of a broken disc. The software automatically finds the disc center and radius from the scanned data, then performs groove rectification and track separation. Starting from the light intensity curve of the pixels in the image, the groove is modeled and the audio samples are obtained. The equipment for the laboratory includes items for the editing, post-production, playing and recording, digital transfer, monitoring, and processing of audio documents as well as for the visual documentation of the whole preservation and restoration process. In the laboratory is used only open source software (often developed by the consortium), with clear advantages in quality, reliability, flexibility and cost. The selection criteria (of the audio documents to be A/D transferred) adopted trades off between:

1) original carriers in immediate danger of terminal decay;
2) documents in regular demand [4];
3) variety of carriers.

A project task focuses on the creation of preservation copies, destined to archive use, and finalized to the transmission of not only the audio, but also the information that characterizes the document together with the sound message (see Sect. IV). The creation of access copies (see Sect. V) provide means for access and prospective self-sustainability of the archive. Access copies will result from the application of novel audio restoration algorithms, developed on purpose (i.e. for the particular characteristics of the Vicentini audio documents). All the filters will be coded in C/C++ into VST plugins. A first result is described in Sect. V-A. By implementing advanced cultural engineering issues, REVIVAL in 2010 will design and prototype a on-line system for cataloguing and fruition of audio documents, allowing information sharing with the most important audio archives worldwide. Finally, tutorial sessions are regularly planned with permanent personnel working at Arena Foundation so as to ensure the transfer of skills and technologies along the project. This is necessary for further coverage of the results to the whole Arena archive, and for support to preservation programmes external to the Arena Foundation.

III. Arena

In the summer of 1913, to celebrate the centenary of the birth of Giuseppe Verdi, the tenor Giovanni Zenatello and the theatre impresario Ottone Rovato promoted a lyrical festival in Verona: the Arena di Verona (Italy) became the biggest open-air lyrical theatre in the world, a supremacy it still holds today. In 1936 saw the light the *Ente Lirico Arena di Verona* (the autonomous organization for lyrical productions of Arena), which took the festival under its care, until it was transformed into a private law foundation in 1998, the Arena di Verona Foundation. Each summer, the venue keeps attracting an international audience - offering opera enthusiasts the best of a rich repertoire set in one of the most magnificent shells of ancient Roman ampitheatres.

A. Fondo Vicentini

In 2001 the heirs of Dr. Mario Vicentini donated to Arena Foundation a large sound archive consisting

Fig. 1. Some tape recorders in Vicentini archive.

Fig. 2. Some audio documents stored in Vicentini archive.

of tens thousands of audio documents stored on different carriers, hundreds of pieces of equipment for playback and recording (wire, cylinder and magnetic tape recorders, phonographs), bibliographic publications (including monographs and all issues of more than sixty music journals from the 1940's to 1999). Along with a history of the recording techniques documented through the carriers (from steel wire and wax cylinders to digital magnetic tapes), the archive traces the evolution of a composite genre such as opera, with an impressive collection of live and studio recordings (smaller sections of the archive consist of symphonic music and other classical genres). The estimated value of the archive is $2,300,000$ Euros. Figure 1 shows a small example of the equipments preserved by Vicentini. Figure 2 shows some audio documents stored in archive.

IV. REVIVAL PRESERVATION PROTOCOL

A reconnaissance on the most significant positions of the debate evolved since the Seventies inside the archivist community on historical faithfulness of the active preservation points out at least three different perspectives [5].

William Storm [6] individuates two types of re-recording which are suitable from the archival point of view: 1) the sound preservation of audio history, and 2) the sound preservation of an artist. The first type of re-recording (Type I) represents a level of reproduction defined as the perpetuation of the sound of an original recording as it was initially reproduced and heard by the people of the era. The second type of re-recording (Type II) was presented by Storm as a more ambitious research objective: it is characterized by the use of a different playback equipment than the original one, with the intent of obtaining the live sound of original performers, transcending the limits of a historically faithful reproduction of the recording.

Schüller [1] and (cited in) [7] points directly towards defining a procedure which guarantees the re-recording of the signals the best quality by limiting the audio processing to the minimum. He goes on to an accurate investigation of signal alterations, which he classifies in two categories: (1) intentional and (2) unintentional. The former includes recording, equalization, and noise reduction systems, while the latter is further divided into two groups: (i) caused by the imperfection of the recording technique of the time (distortions), and (ii) caused by misalignment of the recording equipment (wrong speed, deviation from the vertical cutting angle in cylinders or misalignment of the recording in magnetic tapes). The choice whether or not to compensate for these alterations reveals different re-recording strategies: (A) the recording as it was heard in its time (Storm's Audio History Type I); (B) the recording as it was produced, precisely equalized for intentional recording equalizations (1), compensated for eventual errors caused by misaligned recording equipment (2ii) and replayed on modern equipment to minimize replay distortions; (C) the recording as produced, with additional compensation for recording imperfections (2i).

George Brock-Nannestad [8] examines the re-recording of acoustic phonographic recordings (pre-1925). In order to have scientific value, the re-recording work requires a complete integration between the historical-critical knowledge which is external to the signal and the objective knowledge which can be inferred by examining the carrier and the degradations highlighted by the analysis of the signal.

Starting from these positions, REVIVAL defines the preservation copy as a digital data set that groups the information carried by the audio document, considered as an artifact. It aims at preserving the documentary unity, and its bibliographic equivalent is the facsimile or the diplomatic copy. Signal processing techniques are allowed only when they are finalized to the carrier restoration. Differing from the Schüller position, it is our belief that – in a preservation copy – only the intentional alterations (1) should be compensated (correct equalization of the re-recording system and decoding of any possible intentional signal processing interventions). On the contrary, all the unintentional alterations (also those caused by misalignments of the recording equipment) could be compensated only at the access copy level: these imperfections/distortions must be preserved because they witness the history of the audio document transmission.

The A/D transfer process should represent the original document characteristics as it arrived to us. According to the indications of the international archive community [2], [4]: 1) the recording is transferred from the original carrier; 2) if necessary, the carrier is cleaned and restored so as to repair any climactic degradations which may compromise the quality of the signal; 3) re-recording equipment is chosen among the current professional equipment available in order not to introduce further distortions, respecting the original mechanical analogies; 4) the sampling frequency and bit rate must be chosen in respect of the archival sound record standard (at least, 48 kHz / 24 bit, following the slogan: *the worse the signal, the higher the resolution*); 5) the digital audio file format should support high resolution, and should be transparent with simple coding schemes, without data reduction.

The process of active preservation, produces a particularly large and various set of digital documents, which are made up of the audio signal, the metadata and the contextual information (the term metadata indicates content-dependent information that can be automatically extracted by the audio signal; contextual information indicates the additional content-independent information).

In order to preserve the documentary unity it is therefore necessary to digitize contextual information, which is included in the original document and the metadata which comes out from the transfer process: the information written on edition containers, labels and other attachments should be stored with the preservation copy as static images, as well as the photos of clearly visible carrier corruptions. A video of the carrier playing – synchronized with the audio signal – ensures the preservation of the information on the carrier (physical conditions, presence of intentional alterations, corruptions, graphical signs). The video file should be stored with the preservation copy. The selected resolution and the compression factor must at least allow to locate the signs and corruptions of the support. The resolution of 320x240 pixels and a DV JPEG compression, with no more of 65% of quality is considered a satisfactory trade-off by REVIVAL. In this way, REVIVAL uses metadata extraction methods to describe the carrier corruptions presented in [9].

Finally, a descriptive card (consistent with the database of the audio archive) with all the information on the A/D equipment used and the original document, will be stored in the preservative copy. In the case of phonographic discs the playback speed of the equipment (phonograph or electric turntable) is compared with Photos of GHOSTS [3], which is a useful reference because it is not subject to mechanical variations. The results are able to explain the pitch variations of the signal: this is a necessary metadata in the audio restoration task.

V. Access copy

Different approaches can be adopted in a combinated way with audio restoration algorithms, in accordance with the final purposes of the access copy:

1) Documental approach: in this case, the de-noise algorithms only concern the cases in which the internal evidence of the degradation is unquestionable, without going beyond the technological level of that time.

2) Aesthetical approach: it pursues a sound quality that matches the actual public's expectations (for new commercial editions).

3) Sociological approach: it has the purpose of obtaining a historical reconstruction of the recording as it was listened to at the time (see Storm, Type I).

4) Reconstructive approach: it has the objective to preserve the intention of the author (see Storm, Type II).

In this project, the authors are developing innovative algorithms, specifically dedicated to the restoration of the musical recordings stored in Vicentini archive, able to offer satisfying solutions to the problems connected with the time-varying feature peculiar to musical signals. The first algorithm developed, dedicated to the restoration of audio signal re-recorded from shellac discs, is described below.

A. A restoration algorithm dedicated to shellac discs

The most widespread techniques (Short Time Spectral Attenuation, STSA) employ a signal analysis through the Short-Time Fourier Transform (which is calculated on

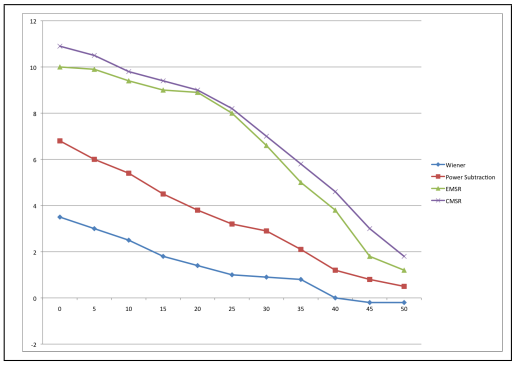


Fig. 3. Gain trend introduced by the filters in the frequency domain at the varying of the input SNR (SNR$_{out}$-SNR$_{in}$ vs. SNR$_{in}$ in dB). The best gain of the CMSR filter can be observed for all the SNR$_{in}$.

small partially overlapped portions of the signal) and can be considered as a non-stationary adaptation of the Wiener filter in the frequency domain. The time-varying attenuation applied to each channel is calculated through a determined *suppression rule*, which has the purpose to produce an estimate (for each channel) of the noise power. A typical suppression rule is based on the Wiener filter [10]: usually the mistake made by this procedure in retrieving the original sound spectrum has an audible effect, since the difference between the spectral densities can give a negative result at some frequencies. Should we decide to arbitrarily force the negative results to zero, in the final signal there will be a disturbance, constituted of numerous random frequency pseudo-sinusoids, which start and finish in a rapid succession, generating what in literature is known as *musical noise*.

More elaborated suppression rules depend on both the relative signal and on *a priori* knowledge of the corrupted signal, that is to say, on *a priori* knowledge of the probability distribution of the under-band signals [10]. A substantial progress was made with the solution carried out in Ephraim and Malah [11], aims at minimizing the mean square error (MSE) in the estimation of the spectral components (Fourier coefficients) of the musical signal. The gain applied by the filter to each spectral component doesn't depend on the simple Signal to Noise Ratio (Wiener Filter), but it is in relation with the two parameters Y_{prio} (SNR calculated taking into account the information of the preceding frame) and Y_{post} (SNR calculated taking into account the information of the current frame). A parameter (α) controls the balance between the current frame information and that of the preceding one. By varying this parameter, the filter smoothing effect can be regulated. Y_{prio} has less variance than Y_{post}: in this way, it is less probable that a musical noise occurs (see [11] for details).

Unfortunately, in the case of shellac discs an optimal value of α does not exist, as it should be time-varying (because of the cycle-stationary characteristics of the disc surface corruptions). Considering this, the authors have developed a new suppression rule (Canazza-Mian Suppression Rule, CMSR), based on the idea of using a *punctual* suppression without memory (Wiener like) in the case of a null estimate of Y_{post}. The pseudo-code is:

IF $Y_{post}\ (k,p) > 0$
$\alpha = 0.98$
ELSE
$\alpha = 0$
END

The experiments carried out, confirm that the filter performs very well during the listening phase, with a noise removal decidedly better than the classic EMSR and the prerogative of not introducing musical noise, at least for $SNR \in [10 \div 20]$ dB (a typical value in the audio signal re-recorded from the shellac discs stored in Vicentini archive). Furthermore, the behavior in the transients is similar of the EMSR filter, without having the perceptual impression of a processing "low-pass filter" like. Figure 3 shows the gain trend introduced by CMSR in comparison with some standard filters (Wiener, Power Subtraction [10], EMSR) at the varying of the noisy signal SNR, considering 20 shellac disc recorded from 1910 to 1930[1]. The term gain indicates the difference between the de-noised signal SNR and the input signal SNR. For all input SNR, the CMSR has a good performance.

The audio alignment method described in [9] is used to compare the differences and similarities between the audio signals stored in the preservative copy and the restored signals. In this way, possible errors occurred in the audio restoration processing are automatically pointed out.

B. Experimental results

As a case study, we selected the double-sided 78 rpm 10 shellac disc Brunswick 58073, recorded in New York on February the 23rd 1928. We considered the song: Rosina Gioiosa Trubia, *Sta terra nun fa pi mia* (*This land is not for me*). Figures 4 and 5 show the (Welch) periodograms (about 15 seconds) of the signal, respectively digitalized by means of a turntable and a phonograph, and restored using CMSR. It can be appreciated that the spectrum of the restored version strictly follows the

[1]A characteristic sample of the archive

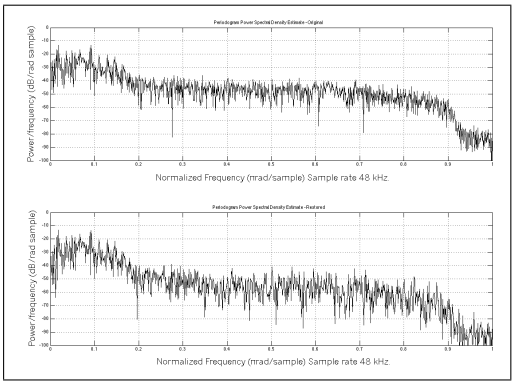


Fig. 4. Periodograms of the audio signals digitized by means of a turntable (top) and restored with CMSR (bottom).

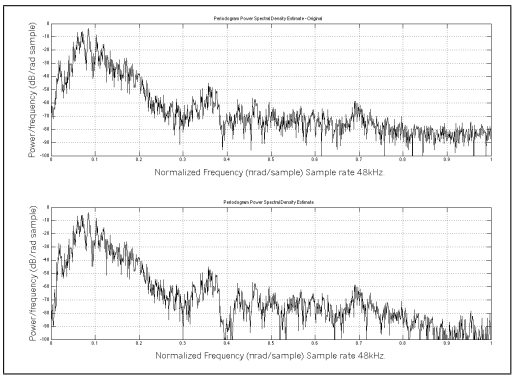


Fig. 5. Periodograms of the audio signals digitized by means of a phonograph (top) and restored with CMSR (bottom).

original one up to frequencies where the noise power density equals that of the musical signal. In addition, differently from what would be obtained with a simple low-pass filter or classical spectral subtraction methods, the restored version "follows" the musical spectrum even in higher frequencies. This aspect is perceptually important and it is appreciated by experienced listeners.

In particular, to use the CMSR to reduce the noise of an A/D transfer by phonograph could be a good approach to obtain a historical reconstruction of the recording as it was listened to at the time.

VI. Conclusion

With its tens of thousands of audio documents, some of which extremely rare or even unique, the Vicentini archive of the Arena di Verona Foundation has potential to gain highest cultural interest worldwide. In spite of this potential, a large part of the archive is at risk of vanishing due to the wear of time. The paper presented the REVIVAL (REstoration of the Vicentini archive In Verona and its accessibility as an Audio e-Library) project, which objective is to establish a HW/SW platform for i) preserving, ii) restoring, and iii) cataloguing the audio documents stored in the archive. Besides its musicological impact, the project aims at fertilizing synergies with national and international public and private bodies, interested in supporting as well as taking advantage of large-scale preservation programs.

Acknowledgment

This work was partially supported by the Foundation Arena di Verona and the University of Verona (Italy) within the framework of the joint cooperation project REVIVAL, 2009-2010.

References

[1] D. Schüller, "Preserving the facts for the future: Principles and practices for the transfer of analog audio documents into the digital domain," *Journal of Audio Engineering Society*, vol. 49, no. 7-8, pp. 618–621, 2001.

[2] AES-11id-2006, *AES Information document for Preservation of audio recordings – Extended term storage environment for multiple media archives*. AES, 2006.

[3] S. Canazza, G. Ferrin, and L. Snidaro, "Photos of ghosts: Photos of grooves and holes, supporting tracks separation," in *Proceedings of XVII CIM*, LaBiennale, Ed., 2008, pp. 171–176.

[4] IASA-TC 03, *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*, D. Schüller, Ed. IASA Technical Committee, 2005.

[5] A. Orcalli, "On the methodologies of audio restoration," *Journal of New Music Research*, vol. 30, no. 4, pp. 307–322, 2001.

[6] W. Storm, "The establishment of international re-recording standards," *Phonographic Bulletin*, vol. 27, pp. 5–12, 1980.

[7] G. Boston, *Safeguarding the Documentary Heritage. A guide to Standards, Recommended Practices and Reference Literature Related to the Preservation of Documents of all kinds*. UNESCO, 1988.

[8] G. Brock-Nannestad, "The objective basis for the production of high quality transfers from pre-1925 sound recordings," in *AES Preprint n°4610 Audio Engineering Society 103rd Convention*, New York, 1997, pp. 26–29.

[9] N. Orio, L. Snidaro, and S. Canazza, "Semi-automatic metadata extraction from shellac and vinyl discs," in *Proceedings of AXMEDIS*, 2008, pp. 38–45.

[10] S. Godsill and P. J. W. Rayner, *Digital Audio Restoration*. Springer, 1998.

[11] D. M. Y. Ephraim, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, and Signal Process*, vol. 32, no. 6, pp. 1109–1121, 1984.

Timbre Similarity Search with Metric Data Structures

Francisco Costa
Msd student
FCT – Universidade Nova de Lisboa
Portugal
fmc17606@fct.unl.pt

Fernanda Barbosa
CITI / DI
FCT – Universidade Nova de Lisboa
Portugal
fb@di.fct.unl.pt

Abstract—Similarity search is essential in music collections, and involves finding all the music documents in a collection, which are similar to a desired music, based on some distance measure. Comparing the desired music to all the music in a large collection is prohibitively slow. If music can be placed in a metric space, search can be sped up by using a metric data structure. In this work, we evaluate the performance of the timbre range query in music collections with 6 metric data structures (LAESA, GNAT, VP-Tree, HDSAT2, LC and RLC) in 2 metric spaces. The similarity measures used are the city-block and the Euclidean distances. The experimental results show that all the metric data structures speeds the search operation, i.e. the number of distance computed in each search process is small when compares to the number of objects in the database. Moreover, the LAESA data structure has the best performance in the two metric spaces used, but the RLC data structure is the only data structure that never degraded its performance and competes with the other metric data structures, in all the experimental cases.

I. INTRODUCTION

With the rapid increase in the use of digital technology, large amounts of music collections will soon be accumulated, like iTunes, emusic and amazon.com. Music browsing are based on the concept of similarity between musical documents, i.e. searching music documents which are very similar or close to a given music document. There are different dimensions to take care in music similarity search [1]. Some of them are melody, harmony, timbre, rhythm and orchestration. But in all of these dimensions, each music document is represented as a vector of numeric properties (features) extracted from the contend-based music.

Currently there are many works related to music similarity search. In these works, the similarity criterion may be based in the melodic dimension [2, 3, 4, 5], in the timbre dimension [6, 7, 8] or in the rhythm dimension [9, 10]. In the most of them the similar searching for a given music document leads to an exhaustive search in the music collection, so the response time will be very long and the search will became ineffective. For this reason, it is necessary to introduce new techniques that can deal with this problem effectively.

The similarity between two music documents is associated with a function, which measures the distance between their respective feature vectors. Some measures used in the music similarity are: Euclidean distance, city-block distance, global edit distance [2, 4], earth mover's distance [11], dynamic time warping and proportional transportation distance [3]. Some of these functions are metric, as Euclidean, city-block and global edit distances. When this function is metric, the set of music documents defines a metric space.

In order to have efficient similar searching in metric spaces, several metric data structures have been proposed [12,13]. These data structures partition the database based on distances between a set of selected objects and the remaining objects. Space partitions seek to minimize the exhaustive search, i.e. at search time, some subsets are discarded and others are exhaustively searched. The distance-based indexing method may be pivot based or cluster based [12]. Some of the data structures using the pivot-based method are the VP-Tree [14] and the MVP-Tree [15]. There are variants of the pivot-based method, used in LAESA [16]. Some of the data structures using cluster-based method are the GNAT [17], the HDSAT [18], the LC [19] and the RLC [20]. The VP-Tree was used in melody search [2] and the RLC metric data structure was already evaluated in different application domains [21, 22, 23].

Our main goal is to evaluate the use and the efficiency of similar searching with metric data structures in music collections. In this work, we address the problem of timbre similarity search in music collections using the metric data structures. This work involves two similarity criterions: the Euclidean distance and the city-block distance. And comprises 6 metric data structures: VP-Tree, LAESA, GNAT, HDSAT2, LC and RLC.

The rest of the paper is structured as follows. In Section II, we recall some basic notions on similarity search in metric spaces. Section III is devoted to the characterization of the metric spaces over music collections. Then Section IV reports the experimental results of the timbre similarity search over music collections. Conclusions and future work are drawn in Section V.

II. SIMILARITY SEARCH IN METRIC SPACES

A metric space is a pair (U,d), where U is a set of objects, called the universe, and d: U x U $\Rightarrow \Re^+_o$ is a function, called distance, that satisfies the three following properties:

- Strict positiveness: $d(x,y) \geq 0$ and $d(x,y) = 0 \Leftrightarrow x = y$;
- Symmetry : $d(x,y) = d(y,x)$;

- Triangle inequality: $d(x,y) \leq d(x,z) + d(z,y)$.

A database over a metric space (U,d) is a finite set $B \subseteq U$.

The output produced by similarity search is a set of objects in the database, which are similar to a given query object q, based on some distance measure d. The similarity search can be range query or k-nearest neighbor.

The result of the range query is the set of objects in the database whose distance to a given object does not exceed a certain amount. Formally, given a database B over a metric space (U,d), a query point $q \in U$, and a query radius $r \in \Re^+$, the answer to the range query (q,r) is the set X, defined in (1).

$$X = \{x \in B \mid d(x,q) \leq r\} \quad (1)$$

The result of the k-nearest neighbor search is the set of closest objects to a given object in the database, where the cardinality of the set is k. Formally, given a database B over a metric space (U,d), a query point $q \in U$, and a positive integer k, the answer to the k-nearest neighbor search $(q)_k$ is the set X with $|X| = k$, defined in (2).

$$X = \{x \in B \mid \forall u \in B\text{-}X, d(x,q) \leq d(u,q)\} \quad (2)$$

Metric data structures seek to minimize the number of distance computations performed in similarity search. During the computation of similarity searching in a database over a metric space (U,d), triangle inequality and symmetry are used to discard some elements of the database without computing the associated distance to the query object. Given a query element q and a radius r, an element x may be left out without the evaluation of d(q,x), if there is an object o where $|d(q,o) - d(x,o)| > r$. In these cases, it is not necessary to compute d(q,x) since we know that $d(q,x) > r$, based on the triangle inequality.

It is important to remark that similarity search is hard to compute in high dimension metric spaces. The calculation in real metric spaces is still an open problem [13]. But it is well know that the metric space dimension grows with the mean and decreases with the variance.

III. MUSIC METRIC SPACE

Our experiments involve only one music collection[1], where each music document is in audio form with sampling rate of 11025 Hz. The collection contains 250 music documents with different music genres [24]: alternative, alternative rock, ambient, dance, electronic, goth and dark era, gothic rock, gothic metal, hard rock, instrumental, metal, new age, nu metal, punk, punk rock, rock, soundtrack, symphonic metal and trance.

Our goal is the timbre similarity search, so we need to characterize the timbre dimension in each music document and use a metric that measures the timbre similarity. In our experiments, we used two metric spaces, i.e. we used two measures: Euclidean and city-block distances.

[1] We do not find any online music collection with songs in audio format. So we create a personal music collection.

A. *Timbre Representation*

The perception of the timbre is related to the structure of the spectrum of a signal, i.e. by its representation in the frequency and temporal domains. The spectrogram and other alternative time-frequency representation are not suitable content descriptors, because of their high dimensionality [1]. A set of descriptors that have been extensively used in music information retrieval are the Mel-Frequency Cepstral Coefficients (MFCCs) [25].

For each music document, we compute an audio signature. The audio signature process was based on the process used by Alfie Tan Kok Leong [26] (see Figure 1). The steps involved in creating an audio signature are:

- Dividing the audio signal into frames with 25.6 milliseconds with an overlapping of 15.6 milliseconds;
- Computing the MFCC for each frame. In our process only the first 13 coefficients are used, because the addition of more MFCC coefficients does not improve performance, as we can see in [26];
- Computing K-means clustering with 1 cluster specified, i.e. the coefficients mean of the music frames, in order to discover the song structure and have a short description or reduction in the information of the waveform.

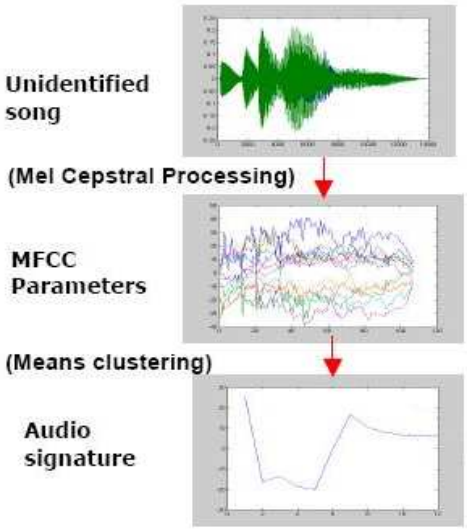


Figure 1. The audio signature process (adapted of [26])

So, each music document has an associate feature vector with size 13.

B. *Timbre Similarity*

In our experiment, the similarity between two music documents is based on the similarity between the associated feature vectors, which is computed with the Euclidean distance and with the city-block distance.

Let $<s_1, \ldots, s_{13}>$ and $<t_1, \ldots, t_{13}>$ be feature vectors associated with two music documents S and T, respectively.

The Euclidean distance between S and T, denoted by ED(S,T), is defined in (3).

$$ED(S,T) = \sqrt{\sum_{i=1..13} (s_i - t_i)^2} \quad (3)$$

The city-block distance between S and T, denoted by CBD(S,T), is defined in (4).

$$CBD(S,T) = \sum_{i=1..13} |s_i - t_i| \quad (4)$$

These measures were evaluated in [26], and had satisfactory results according to the music similarity.

In order to study the metric spaces, we computed the histogram of distances between any music documents of the database, using the two measures. In Table I, we presented the mean and the variance of the histogram of distances.

TABLE I. MEAN AND VARIANCE OF DISTANCES

	city-block distance	**Euclidean distance**
Mean	5.791928	2.413186
Variance	5.075894	1.244004
Mean/ Variance	1.14	1.94

An immediate conclusion is that our metric spaces have different dimensions. The dimension is highest for the metric space with Euclidean distance, where the quotient between the mean and the variance is 1.94. The metric space with city-block distance has lowest dimension.

IV. EVALUATION OF METRIC DATA STRUCTURES

The goal of this section is to understand how range queries with metric data structures behave in the music collection, when the similarity criterion is based on timbre dimension (music metric spaces define in Section III).

For the music collection, four files were generated. The smallest is the set of query music documents and the other three are random permutations of the collection. The use of three equal sets lies in the fact that the final shape of some data structures depends on the order in which the objects occur in the input of the construction algorithm. The size of the queries set is 25% of the database size (63 music documents). For each artist/album, we selected 25% of his music documents, in order to have a query set representative of the database, i.e. all the artists, albums and genres are presented in this set.

For each set associated to the database, we submitted the set of query music documents to range query with different query radii in the two metric spaces. The query radii selected for each metric space were equal to 40%, 60% and 80% of the metric space's mean distance. So for the metric space with Euclidean distance, the query radii were 0.964, 1.5 and 1.928. And for the metric space with city-block distance, the query radii were 2.316, 3.5 and 4.632.

In Tables II and III, we presented the average number of music documents retrieved in range queries, and the associated percentage of the database size, for each pair of metric space and query radius.

TABLE II. AVERAGE NUMBER OF MUSIC DOCUMENTS AND PERCENTAGE OF THE DATABASE RETRIEVED WITH THE CITY-BLOCK DISTANCE

city-block distance					
Query Radius 2.316		**Query Radius 3.5**		**Query Radius 4.632**	
Num	Percent	Num	Percent	Num	Percent
7.7	3.07%	38.8	15.51%	92.1	36.84%

TABLE III. AVERAGE NUMBER OF MUSIC DOCUMENTS AND PERCENTAGE OF THE DATABASE RETRIEVED WITH THE EUCLIDEAN DISTANCE

Euclidian distance					
Query Radius 0.964		**Query Radius 1.5**		**Query Radius 1.928**	
Num	Percent	Num	Percent	Num	Percent
13.3	5.33%	59.2	23.66%	102.9	41.18%

In each experimental case (a given metric space and a given query radius), we computed the average number of distance computations done for each music query. So, the results presented are the mean of results obtained to query the three sets associated to the database.

A. Parameterization of Metric Data Structures

In our experiments, we used 9 metric data structures: LAESA, VP-Tree, DSAT, HDSAT1, HDSAT2, LC with fixed cluster size, LC with fixed radius, GNAT and RLC. The metric data structures were parameterized in order to obtain the best results for the music collection. But we've decide to present only the results of 6 metric data structures, so we selected only one data structure for each "representation", which had best results. In the group DSAT, HDSAT1 and HDSAT2, we choose HDSAT2, and in the group LC, we choose LC with fixed cluster size. The parameterization used in each selected data structure is:

- LAESA – Linear Approximating and Eliminating Search Algorithm, with 19 and 11 pivots for the metric spaces with city-block and Euclidean distances, respectively;
- VP-Tree – Vantage Point Tree, which does not have variants or parameters;
- HDSAT2 – Hybrid Dynamic Spatial Approximation Tree2, with arity 5 and 8 for the metric spaces with city-block and Euclidean distances, respectively;
- LC – List of Clusters, with fixed cluster size 8 and 38 for the metric spaces with city-block and Euclidean distances, respectively;
- GNAT – Geometric Near-neighbor Access Tree, with degree 10;

- RLC – Recursive List of Clusters, with array capacity 30 and radius 6.1 for the metric space with city-block distance, and with array capacity 24 and radius 1.86 for the metric space with Euclidean distance.

B. Experimental Results

Figures 2, 3 and 4 depict the average number of distances computed to query the music collections with the three query radii in the metric space with the city-block distance.

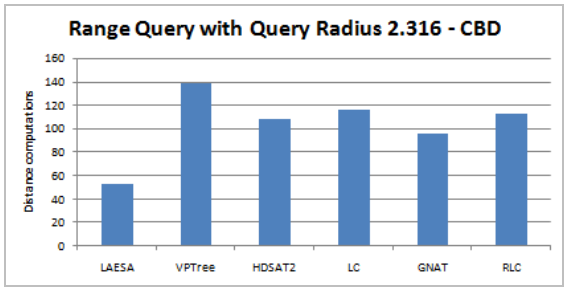


Figure 2. The average number of distance computed to querying the database with city-block distance (query radius 2.316)

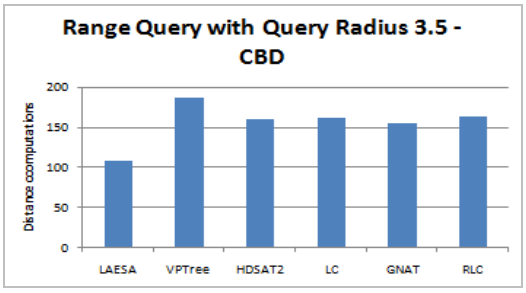


Figure 3. The average number of distance computed to querying the database with city-block distance (query radius 3.5)

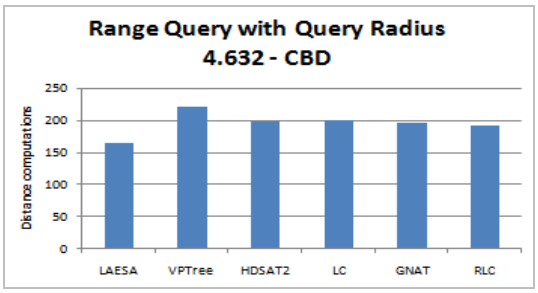


Figure 4. The average number of distance computed to querying the database with city-block distance (query radius 4.632)

Figures 5, 6 and 7 depict the average number of distances computed to query the music collections with the three query radii in the metric space with the Euclidean distance.

In an exhaustive searching, for each query we need to compare the music document query with the all music documents in the database. In our experimental, for all the metric data structures the average number of distance computations by each query is small when comparing with the database size. The corresponding percentage of the database size, for each range query, is shown in Table IV.

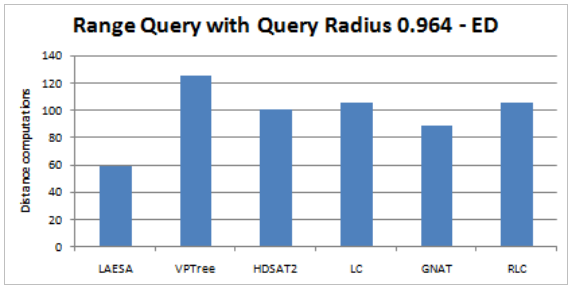


Figure 5. The average number of distance computed to querying the database with Euclidean distance (query radius 0.964)

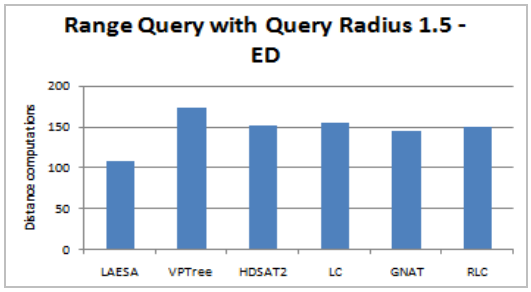


Figure 6. The average number of distance computed to querying the database with Euclidean distance (query radius 1.5)

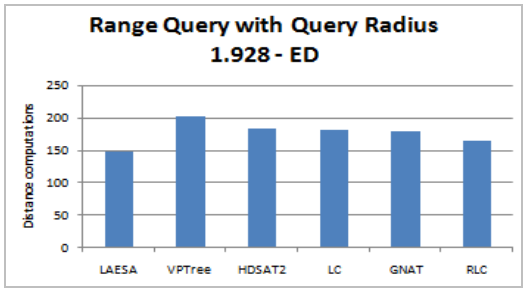


Figure 7. The average number of distance computed to querying the database with Euclidean distance (query radius 1.928)

TABLE IV. THE PERCENTAGE OF DISTANCE COMPUTATIONS ACCORDING TO THE DATABASE SIZE

	City-block distance			Euclidean distance		
Query Radius	**2.316**	**3.5**	**4.632**	**0.964**	**1.5**	**1.928**
LAESA	21%	43%	66%	24%	44%	59%
VP-Tree	55%	75%	88%	50%	70%	81%
HDSAT2	43%	64%	79%	40%	61%	73%
LC Fixed Cluster Size	46%	65%	80%	42%	62%	73%
GNAT	38%	62%	78%	36%	58%	71%
RLC	45%	66%	77%	42%	60%	66%

These results confirm that a lot of music documents are discarded without computing the associated distance to the music document query.

We can observe that the best results were obtained in LAESA data structure, for all the query radii in the two metric

spaces. And the worse results were obtained in VP-Tree data structure. Excluding LAESA, all data structures were very competitive, but it is important to remark two situations in these data structures:

- The GNAT had the best results with the smallest query radius in the two metric spaces;
- The RLC was the only data structure that never degraded its performance when the query radius increased, i.e. with the two smallest query radii, the RLC competes with the other data structures, but in biggest query radii, RLC is the best data structures.

All the metric data structures had best results in the metric space with the Euclidean distance, which is the highest dimension space.

V. CONCLUSIONS AND FUTURE WORK

The need to speed the music browsing, lead us to evaluate the performance of range queries with metric data structure in music similar search.

The size of our music collection is not very representative to the real size of music collections. So our next priority is to make this evaluation in large music collections, and we expected to have better results in these collections.

With respect to our metric spaces (timbre representation, and city-block and Euclidean distances), we know that there are many other ways to search for music similarity. So, we have an ongoing work related to search for melodic similarity. We also pretend to evaluate the metric data structures in the rhythm similarity search.

With respect to the efficiency of the range queries with metric data structures, the results leaves us to conclude that the metric data structures speed the range query in the two metric spaces used. This conclusion is based on the observation that a lot of music documents are discarded without computing the associated distance to the music query. The LAESA data structure has the best performance in the two metric spaces, but the RLC data structure is the only data structure that never degraded its performance and competes with the other metric data structures, in all the experimental cases.

REFERENCES

[1] N. Orio, *Music Retrieval: A Tutorial and Review*, Now Publishers Inc, 2006.

[2] M. Skalak, J. Han, and B. Pardo, "Speeding Melody Search With Vantage Point Trees in *Proc. of International Society for Music Information retrieval (ISMIR)*, 2008.

[3] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R. Van Oostrum, "Using Transportation Distances for Measuring Melodic Similarity", in *Proc. of International Society for Music Information retrieval (ISMIR)*, 2003.

[4] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis, "A Comparative Evaluation of Search Techniques for Query-by-Humming Using the MUSART Testbed", *Journal of the American Society for Information Science and Technology, pp. 687-701,* 2007.

[5] B. Pardo, and W. Birmingham, "Encoding Timing Information for Musical Query Matching", in *Proc. of International Society for Music Information retrieval (ISMIR)*, 2002.

[6] B. Logan, and A. Salomon, "A music similarity function based on signal analysis", *Proc. of International Conference on Multimedia and Expo (ICME)*, August ,2001.

[7] J.-J. Aucouturirer, and F. Pachet, "Music Similarity Measures: What's the Use?", in *Proc. of International Society for Music Information retrieval (ISMIR)*, 2002.

[8] E. Pampalk, "A Matlab toolbox to compute music similarity from audio", in *Proc. of International Society for Music Information retrieval (ISMIR)*, 2004.

[9] J. Foote, M. Cooper, and U. Nam, "Audio Retrieval by Rhythmic Similarity." in *Proc. of International Society for Music Information retrieval (ISMIR)*, 2002.

[10] J. Foote, and U. Nam, "The Beat Spectrum: A New Approach to Rhythm Analysis", *Proc. of International Conference on Multimedia and Expo (ICME)*, 2001.

[11] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval", *Tech. Report*, Stanford University, 1998

[12] H. Samet, *Foundations of Multidimensional and Metric Data Structures*, Morgan Kaufmann Publishers, San Francisco, USA, 2006.

[13] E. Chávez, G. Navarro, R. Baeza-Yates, and J. Marroquín, Searching in metric spaces, *ACM Computing Surveys*, 33, 3, pp. 273-321, 2001.

[14] P. Yianilos, "Data structures and algorithms for nearest neighbor search in general metric spaces", in *Proc. of the 4th Annual SIAM Symposium on Discrete Algorithms*. ACM Press, USA, pp. 311-321, 1993.

[15] T. Bozkaya, and M. Ozsoyoglu, "Distance-based indexing for high-dimensional metric spaces", in *Proc. of the SIGMOD International Conference on Management of Data (SIGMOD'97)*. ACM Press, New York, UK, pp. 357-368, 1997.

[16] M. Micó, J. Oncina, and E. Vidal, A new version of the nearest-neighbour approximating and eliminating search algorithm. *Pattern Recognition Letters*, 15, 1, pp. 9-17, 1994.

[17] S. Brin, "Near neighbor search in large metric spaces" in *Proc.* of *21st International Conference on Very Large Data Bases (VLDB'95)*. Morgan Kaufmann Publishers, Zurich, Switzerland, pp. 574-584, 1995.

[18] D. Arroyuelo, F. Muñoz, G. Navarro, and N. Reyes, "Memory-adaptative dynamic spatial approximation trees", in *Proc. Of the 10th International Symposium on String Processing and Information Retrieval (SPIRE)*, 2003.

[19] E. Chávez, and G. Navarro, A compact space decomposition for effective metric indexing. *Pattern Recognition Letters*, 26, 9, pp. 1363-1376, 2005.

[20] M. Mamede, "Recursive Lists of Clusters: a dynamic data structure for range queries in metric spaces", in *Proc. of the 20th International Symposium on Computer and Information Sciences (ISCIS 2005)*. Springer-Verlag, Berlin, Germany, pp. 843-853, 2005.

[21] M. Mamede, and F. Barbosa, "Range Queries in Natural Language Dictionaries with Recursive Lists of Clusters", in *Proc. of the 22th International Symposium on Computer and Information Sciences (ISCIS 2007)*, IEEE Xplore (doi:10.1109/ISCIS.2007.4456857), 2007.

[22] F. Barbosa, "Similarity-based retrieval in high dimensional data with Recursive Lists of Clusters: a study case with Natural Language Dictionaries", in *Proc. of the International Conference on Information management and engineering (ICIME 2009)*, IEEE Computer Society (ISBN: 978-1-4244-3774-0), 2009.

[23] F. Barbosa, and A. Rodrigues, "Range Queries over Trajectory Data with Recursive List of Clusters: a case study with Hurricanes data", in *Proc. of Geographic Information Science Research UK (GISRUK 2009)*, UK, 2009.

[24] Lastfm, Music community website, www.lastfm.com

[25] B. Logan, "Mel Frequency Cepstral Coeffcients for Music Modeling", in *Proc. of International Society for Music Information retrieval (ISMIR)*, 2000.

[26] Alfie Tan Kok Leong**,** *A Music Identification System Based on Audio Content Similarity*, Doctoral Thesis, University of Queensland, Queensland, Australia, 2003.

A location-tracking interface for ethnomusicological collections

Michela Magas
Goldsmiths Digital Studios
University of London
London SE14 6NW, UK
Email: map02mm@gold.ac.uk

Polina Proutskova
Department of Computing, Goldsmiths
University of London
London SE14 6NW, UK
Email: mas02pp@gold.ac.uk

Annotation in ethnomusicological archives is often incomplete, and searches involving artist's or composer's name often do not apply. We propose a novel system for intuitive exploration of ethnomusicological recordings which combines a geographical visual interface with ethnomusicological annotation, similarity matching and location tracking.

I. Introduction

Music is the best way to approach a culture and its people: appreciating ethnomusicological recordings may help establish the understanding for combating the ills of ethnocentrism, racism and religious fundamentalism [1],[2]. In this paper we present our research into tools for more efficient exploration of ethnomusicological archives. In the opening section we provide an overview of the current state of ethnomusicological archives; we follow by describing existing systems which can be used as interfaces for ethnomusicological recordings; and we present a novel user interface combining annotation with geographical sampling, audio similarity searches, and geographical tracking of search results. We conclude with an outlook and further research suggestions.

II. The current state of ethnomusicological archives

Ethnomusicological recordings document the diversity of musical repertories and traditions in a systematic way and constitute an important part of cultural heritage. Many endangered or extinct music cultures can only be preserved for humanity in the form of recordings safeguarded in ethnomusicological archives. Dissemination of ethnomusicological recordings made possible a shift during the Twentieth Century from a global understanding of music based on ethnocentrism and a belief that music evolved torwards European harmony, to one based on knowledge of the fundamentals, culture and underlying philosophies of the societies in which the music was created [2]. Anthony Seeger [1],[3] describes the significance of the archives for the field of ethnomusicology and the public groups interested in using the archives' resources.

Europe has a long standing tradition of documenting musical cultures by means of audio (nowadays also audio-visual) recordings. European institutions hold historical recordings of key importance not only for European cultures but also musical traditions outside Europe. The first ethnomusicological audio recording was made in 1900 by Erich M. von Hornbostel, when Siamese (Thai) musicians were giving a performance in Berlin. The Berlin Phonogram Archive was founded in the same year in order to collect, preserve and study music recordings of world cultures. We can still listen to this first recording today, as well as to many other wax cylinder recordings preserved at the Berlin Phonogram Archive.

There are huge repositories of ethnomusicological recordings residing at various institutions in Europe. The National Sound Archive (Traditional Music Section) in London hosts one of the largest ethnomusicological collections in the world, containing recordings from all around the globe. The Berlin Phonogram Archive and the Phonogram Archive in Saint Petersburg (Russia) are also very large, culturally spread collections. Many countries fund national sound/music archives, which are the main documenting bodies of local and national music cultures of the state (Polish National Sound Archive, Bulgarian Academy of Sciences Music Archive, etc.). Fig. 1 shows the amount of ethnomusicological recordings in some of the largest archives of the world.

Though the policies of European institutions usually allow researchers and general public to access recordings in ethnomusicological archives, their holdings are still very much underused and are not sufficiently exposed to potential users. In many cases it is still necessary to come physically to the archive in order to search its collections and to listen to the recordings [4].

A considerable number of recordings in European archives has now been digitized according to world standards (IASA TC-04). DISMARC (http://www.dismarc.org/) was the first European project to expose catalogues of ethnomusicological archives online. It provides a metasearch of ethnomusicological collections to museums, universities, national archives and broadcasting companies from Germany, Sweden, Finland and the UK. The number of participants is steadily growing. We envision further growth of digital content in ethnomusicological archives as well as large-scale collaborative projects like DISMARC and EUROPEANA (European Digital Library, http://www.europeana.eu/) to make huge distributed reposito-

	Items	Hours (incl. commercial)	Original collections
Berlin Phonogram Archive	Over 150,000	18,000	Over 1,000
US Library of Congress, Archive of Folk Culture	-	Over 100,000	Over 4,000
National Sound Archive of the British Library, World and Traditional Music	300,000	-	370
Archives for Traditional Music, Indiana University	128,550	Over 250,000	2,000

Fig. 1. Amount of recordings in selected ethnomusicological archives

ries accessible online.

Ethnomusicological archives in Europe, USA and other parts of the world are now reinventing themselves, and there are many obstacles to overcome, including missing infrastructure for large digital libraries, metadata management, effective search strategies and attractive presentation to the user. We address some of these issues in our research described in this paper and suggest scalable solutions.

III. Interfaces for Ethnomusicological Recordings

The growth of ethnomusicological resources searchable online requires new scalable retrieval strategies as well as their intuitive graphical representations [5]. Ethnomusicological collections constitute a challenge with respect to both metadata and content management. The audio content of these collections is still a less studied terrain in Music Information Retrieval [6],[7]. Today most of the online ethnomusicological archives' catalogues offer bibliographic text searches that may sometimes include categories specific to music (eg. The British Library Sound Archive Catalogue simple search, http://cadensa.bl.uk/cgi-bin/webcat). Finding salient results is hindered by the incompatibility of controlled vocabularies between archives and metadata schemes.

In ethnomusicological archives search and retrieval differ significantly from music archives containing Western repertories: there is usually no composer because many recordings are of traditional music; the artist is in most cases not known to the user; there is no such thing as album title and the song title may be just the lyrics incipit; genre and functional descriptions may exist, but they are most definitely not compatible between cultures. The first descriptor according to which an ethnomusicological collection is usually searched is the cultural origin of the music [8], which often (but not always) coincides with the geographic location where the recording was made.

Alan Lomax was the first to suggest a user interface for an ethnomusicological collection based on geographic location, using recordings from the Cantometrics database [9]. The Cantometrics project, started in the 1960s and pursued by Lomax until the 1990s, aimed at discovering correlations between singing styles and societal traits such as hierarchy structures and gender relations. For that purpose Lomax and his colleagues gathered and analysed around 7,000 recordings from more than 500 cultures [9],[10]. In his Global Jukebox [11] clickable icons on the world map trigger the audio content of the recording made at the respective location. A data similarity search is based on the 37 Cantometrics parameters, and relies on the manual annotation of each track by expert listeners. The results of this search can be shown on the geographical map or used to map out charts showing correlations between patterns of song and patterns of society [11].

Portals such as the National Geographic Nat Geo Music portal (http://worldmusic.nationalgeographic.com/) or Mondomix (http://www.mondomix.com/) provide a fairly comprehensive overview of music from all over the globe, including audio and video searches, though they are limited to commercial releases. Their interfaces rely on a simple combination of text and media, offering the user a list of countries to choose from, with limited data visualization, interactivity or content management, and few incentives for intuitive exploration.

Visualisations of commercial recordings rely mostly on clustering by genre or mapping of artist relationships (eg. [12],[13]), which make them inappropriate for ethnomusicological recordings. With classification by genre, attempts to classify music often fail because of reasons like ambiguities, subjective judgment and marketing interests [14]. Musicovery (http://www.musicovery.com/, originally developed as Music Plasma or Live Plasma by Frederic Vavrille) provides a colourful visualization of tracks according to genre and links between similar artists. It provides the option of selecting and listening to the genre of 'world' music, though the interface relies on standard music interface data (artist and name of track) leaving no room for ethnomusicological descriptions. It provides no information about the geographical or cultural origin of the tracks, or why particular tracks have been visually linked, and presents no opportunity for data management and annotation.

Collaborative database projects, such as The Freesound Project (http://www.freesound.org/) which focuses on environmental sound [15], and Xeno-Canto (http://www.xeno-canto.org/) which focuses on bird song [16], rely on voluntary contributions of recordings, and use the Google Maps geotagging system as the basis for exploration of recordings from around the world. Clicking on a pin provides a link to a list of recordings related to the chosen location, which in turn provide links to audio samples. Recordings can be downloaded under the Creative Commons license and freely studied and explored. Tags and descriptions can be collaboratively added. Even

	C	D	E	F	G	H
1	**continent**	**location / culture**	**singing**	**solo / group**	**male / female**	**instruments**
2	Europe	Georgia				
3	N. Europe	England				
4	N. America	Virginia	singing	solo	mixed	guitar
5	N. America	Pueblo Taos	singing	group	male	drum
6	S. America	Interior Amazon, Jivaro				
7	N. Europe	Hebrides, Scotland	singing	solo group	solo female group mixed	no
8	C. Europe	Asturias	singing	group	group female group male	no
9	Afro-America	Andros, Bahamas island	singing	solo group	male	no
10	E. Europe	Daghestan, Caucasus	singing	group	group female group mixed	no
11	C. Africa	Mbuti Pygmies, Equatorial rainforest	singing	group	mixed	no
12	C. Asia	Kazakhstan	singing	solo	male	lute
13	S. Europe	Trapani, Sicily				
14	E. Europe	Molokan, Central Russia				
15	E. Asia	Ainu				
16	N. America	French Canadian				
17	E. Europe	State Siberian Russian Folk Choir	singing	group	group female group male	no
18	W. Africa	Wolof, Senegal				
19	E. Europe	Rhodope, Bulgaria				

Fig. 2. Manual annotation of the Cantometrics Training Tapes dataset.

though they are not designed specifically for ethnomusicological recordings, these collaborative audio-collecting systems offer greater scope for ethnomusicological research.

Geotagging is also used by the British Library Sound Archive online interactive maps which allow researchers to sample, download and study a variety of recordings with related texts and metadata (http://sounds.bl.uk/Maps.aspx). However, none of the above examples make use of the feature extraction and audio similarity tools made available by the latest developments in MIR research.

IV. Enhanced Geographical User Interface

We present a dynamic user interface which offers a novel audio content-based search for ethnomusicological recordings. It makes use of the latest findings on similarity searches in the field of MIR and adds a visual feature for location-tracking the similarity results, thus offering a carefully designed, intuitive gateway for exploring and searching a music collection of any cultural and geographical scope.

The starting point for the design of this interface is a set of ethnomusicological recordings of world music compiled by Alan Lomax as part of the Cantometrics project. We use a collection of 355 recordings from the Cantometrics Training Tapes dataset [9] containing all varieties of singing and playing styles from more than 50 cultural origins. The nature of the collection suggests a search by geo-location. We propose a multi-layered interface which allows for query triggers from a geographical plane located at a level below the audio similarity search interface. The layering of geo-positioning and similarity matching enables the discovery of interesting anthropological and ethnomusicological relationships.

The Cantometrics recordings can be sampled at their geo-locations before a query is triggered. We manually annotated the tracks with 9 attributes: a capella singing, accompanied singing, purely instrumental, choral singing, solo singing, male singing, female singing and mixed male and female singing (Fig. 2). These are represented by simple, colour-coded symbols at their relative location: male or female singer, solo or group, vocal or instrumental. They provide a novel view of the geographic distribution of singing style patterns. In order to preserve the legibility of the symbols and the layers above them, the design of the World map is a minimal, flattened, monochrome version of the Google Maps graphic (Fig. 3).

Global positioning of the individual symbols is done manually, in absence of any geographical coordinates in Lomax's original annotation. This occasionally presents a problem: some locations belong to a different country than they did at the time when the Cantometrics recordings were collected and some of the countries and people's names have since changed. We have tried to keep faithful to the original descriptions as much as possible, though cannot guarantee absolute accuracy.

On query launch, the recorded sample is matched to other Lomax recordings from around the globe and brought to view in the mHashup similarity interface [17]. Our interface can potentially use any audio similarity search engine that is able of finding best matching fragments of audio files in a database. In this example we use the audioDB database management system which employs Euclidean distance in the normed audio shingle space to retrieve similar tracks [18],[19],[20]. For each audio file we extract 20-band Mel Frequency Cepstral Coefficients (MFCC). These are extracted using a short-time Fourier transform with hop size 100ms (2205 samples), window length 185.76ms (8192 samples), FFT length 16384 samples (2.69Hz frequency bins). Feature extraction is performed using fftExtract.

The similarity interface and the geographical interface create layers which can be independently reduced in size or enlarged as required. Layering of query results over their geographical locators allows for an immediate overview of a variety of elements and parameters related to the recordings. The length of a bar corresponds to the relative length of a recording. The position of the matching audio shingle [17] is aligned with that of the queried one, visually highlighted, and allows for further sampling by direct manipulation. All searches are stored, enabling a return to a previously chosen query pathway and a jumping-off point for new searches.

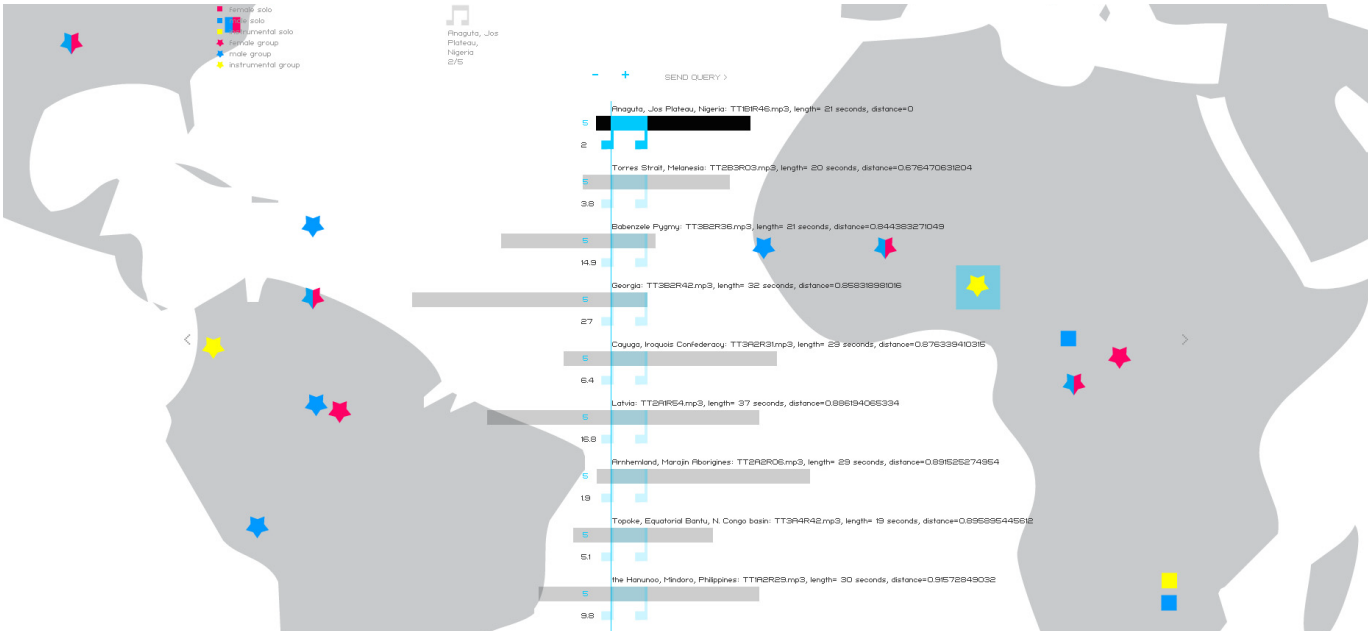


Fig. 3. A location-tracking interface for ethnomusicological recordings. The square location tracker is visible on the right in Nigeria, as the site of the Anaguta, which have been chosen as the initial query for the audio similarity results displayed in the centre. The highlighted query results are displayed on bars corresponding to the relative lengths of tracks, as segments relative to the query length, and are aligned to the query point. Clicking on a query result will play the corresponding audio and launch the location tracker on a trajectory to the corresponding location on the map. Artwork copyright 2009 Michela Magas.

The matched recordings can be sampled individually, with a highlighting device tracking the query trajectory from one part of the globe to another. The location tracker aligns itself with the geographical location of each sampled query result before launching on a trajectory to the next one, creating a web of pathways indicating relationships between cultures in distant locations. Thus looking for recordings similar to highly polyphonic choral singing from Georgia in the Caucasus with passages of yodel in the highest part, the location tracker will travel around the world to Mbuti Pygmies in the Central African rainforest singing complicated interlocked patterns also containing yodel elements, to a choir from Montenegro, to Native Americans from the Iroquois Confederacy, back to Georgia and then to Nigerian pan pipe playing. The location tracker is here intended as a visual aid to the user's understanding of the relationships, connections and distances between ethnomusicological recordings and related cultures.

V. Preliminary User Evaluation

Preliminary user testing was conducted with four male participants of varying ages: 60, 37, 30 and 27. They were encouraged to explore the interface intuitively, without much briefing or explanation. They were then asked four questions: How easy is it to sample a recording? How easy is it to trigger the audio similarity query? How easy is it to sample and compare the results? How easy is it to zoom in and out of layers? The first three questions received consistently high scores (Fig. 4), while the zoom function proved harder to use.

On a touch screen, such as on the Apple iTouch, the zoom function is integral to the device's core set of functionalities and regular users are familiar with the related commands. In such an environment a zoom function is expected by the users and therefore does not require special instructions. The task of implementing a zoom function is more complex in the desktop environment, particularly where it relies on the use of a mouse. Here we rely on a trigger by the zoom + and - symbols for its functioning. These need to be visually flagged very clearly as they are not a common feature of desktop interfaces, and users do not necessarily expect to find them or understand that, in the case of our interface, each layer can be scaled separately. Zooming in and out of layers, while not essential to the operation of our interface, ought to be flagged to the user more prominently as it may be a useful addition, providing more freedom for the exploration of the interface.

Useful suggestions have been provided regarding the sending of queries, such as making the 'Send Query' button relate visually to the latest audio shingle sampled, in order to reinforce the connection and indicate that the said shingle can now be used as query for audio similarity matching. We have responded to this suggestion by making the 'Send Query' button mobile: it now aligns itself with the latest audio shingle sampled, thus indicating the said shingle can be used as a query trigger for further searches.

A participant noted that the audio similarity matching occasionally matches silence to audio query content. This is a

Participant's age:	60	37	30	27	Average score
How easy is it to sample a recording?	8	9	10	9	**9**
How easy is it to trigger the audio similarity query?	9	8	8	10	**9**
How easy is it to sample and compare the results?	9	8	6	9	**8**
How easy is it to zoom in and out of layers?	6	3	3	8	**5**

Fig. 4. Preliminary user evaluation. Results are 1 - 10, where 10 is best.

bug of the current feature extraction process and it is to be further explored. Sampling longer audio shingles of more than 15 seconds may cause a delay in the streaming of audio on certain networks. On these rare occasions, a visual cue ought to be provided to indicate to the user that the audio is about to be streamed. An option to play results simultaneously has been requested and can be provided. Similarly, an option can be provided allowing the users to choose whether they prefer to initialise the audio similarity search according to a match by melody or timbre. The usefulness of the above suggestions will be tested with further user evaluation.

VI. Future work

The interface would benefit from the planned addition of our system for automatic annotation of digital audio recordings [21]. This system automatically extracts basic musical attributes, independently of their cultural origin or musical style, and helps in the addition of consistent metadata to large ethnomusicological repositories, thus enabling applications to large-scale collaborative projects.

Every time a new song is added to the database, feature extraction and automatic annotation would be initiated and reflected in the visual interface by the corresponding symbols and metadata. Automated annotation rarely gives absolutely reliable answers, thus it is no substitute for the work of an archivist or a researcher and must be clearly distinguished from reliable human annotation. Where the archivists and librarians find the metadata inconsistent or incomplete, they ought to be able to edit it by direct manipulation of the interface.

Further musical parameters could be extracted automatically: amount of percussivity, pulse stability, amount of repetition, presence of specific singing production patterns such as yodel, drone or overtone singing. The variety of musical style patterns can then be represented graphically showing their spread around the world.

The enhanced geographic interface could be further automated by including a semi-automatic geographical location component. Geographical coordinates based on the given metadata could allow for the automatic positioning of a recording on the map. Only when the cultural/geographic description cannot be automatically located or is ambiguous, would the manager/archivist be prompted to search for coordinates manually.

The function of the location tracker suggests the idea of 'culture-hopping', which can be further explored and enhanced with the addition of relevant anthropological information. Every time the location tracker lands in a particular location, information about the related culture could be provided: emphasizing, for example, whether this is an historic or current culture, whether it is tribal, linked to agriculture or urban.

Online integration may be very beneficial at this point. People interested in particular recordings might be better informed about their cultural context than the archivists are. Social tagging could be used to counter the errors generated in automatic annotation: an online editorial system would allow registered users to add, edit and negotiate information on catalogue entries, which could then become available to all users. Combined with archivists' expertise and moderation, this approach would enable archives to close gaps in annotation and offer hands-on activities to their user communities.

VII. Conclusion

In expectation of further growth of digital content and online exposure of ethnomusicological collections we have presented a dynamic graphic user interface combining graphic representation of ethnomusicological annotation with an audio content search, and providing a location-tracking visualization of ethnomusicological collections of large scope. Benefits of combining metadata with audio content search and retrieval have often been stressed in MIR literature [22],[23]. Ethnomusicological archives may benefit more than any other content holders from this approach: on the one hand cultural context is especially important in ethnomusicology and audio similarity can only become meaningful in conjunction with cultural information; on the other hand audio similarity search can help to close gaps in missing, poor or inconsistent metadata.

According to preliminary evaluation, the interface fulfilled all of the essential usability requirements and received positive comments. The users were able to explore the collection intuitively, without prior training. Minor amendments and additions suggested will be introduced in future work to enhance the interface's usability further.

Exposing an ethnomusicological collection to a similarity search highlights the huge variance in the music data with respect to cultural origin, musical style and structure, singing production, instruments played as well as recording quality.

This opens possibilities for further research into the extent the similarity search strategy used (both low level audio features and the pattern matching mechanism) corresponds to listeners' perception of similarity, and particularly how this perception varies culturally.

Our combined location-tracking and audio content similarity search can facilitate cross-cultural research in ethnomusicology, as in Victor Grauer's hypothesis about the evolutionary relationship between the pygmy/bushmen singing style and various interlocked pan pipe playing styles [24]. It offers the user the opportunity for creative exploration and unexpected discoveries: a novel way to discover music cultures of the world.

Acknowledgment

This work is part of the OMRAS2 project (http://www.omras2.org) funded by the EPSRC (http://www.epsrc.ac.uk). OMRAS2 aims at creating a comprehensive high level music information retrieval system for musicologists. The authors would like to thank the following OMRAS2 researchers: Michael Casey, Christophe Rhodes, Ben Fields and Tim Crawford.

References

[1] A. Seeger, "The role of sound archives in ethnomusicology today," *Ethnomusicology*, vol. Spring/ Summer, 1986.

[2] R. Reigle, "Humanistic motivations in ethnomusicological recordings," in *Recorded Music - Philosophical and Critical Reflections* (M. Dogantan-Dack, ed.), Middlesex University Press, 2009. with CD.

[3] A. Seeger, "Ethnomusicologists, archives, professional organisations, and the shifting ethics of intellectual property," *Yearbook for Traditional Music*, pp. 87–107, 1996.

[4] P. Proutskova, "Data infrastructure for ethnomusicological archives – current situation and future perspectives", *Journal of International Society of Sound and Audiovisual Archives*, pp. 45–54, july 2008.

[5] P. Proutskova and M. Magas, "Beyond the metadata, new intelligent audio content search for large music collections", presented at the *Unlocking Audio 2* conference, British Library, London, UK, 16-17 Mar., 2009.

[6] G. Tzanetakis, A. Kapur, W. A. Schloss, and M. Wright, "Computational ethnomusicology", *Journal of Interdisciplinary Music Studies*, vol. 1, no. 2, pp. 1–24, 2007.

[7] J. S. Downie, "Music information retrieval," *Annual Review of Information Science and Technology*, vol. 37, pp. 295–340, 2003.

[8] M. Mengel, "ethnoArc, linked european archives for ethnomusicological research," *Tech. Rep. D4*, Berlin Phonogram Archive, February 2007.

[9] A. Lomax, "Cantometrics: An Approach To The Anthropology Of Music", The University of California, 1976. accompanied by 7 cassettes.

[10] A. Lomax, "Folk Song Style and Culture", New Brunswick, NJ: Transaction Books, 1968.

[11] A. Lomax et al, "Global Jukebox", available at http://www.culturalequity.org/video/ce_videos_global_jukebox.jsp.

[12] D. Gleich, M. Rasmussen, K. Lang and L. Zhukov, "The World of Music: SDP layout of high dimensional data", in *Info Vis*, 2005.

[13] O. Goussevskaia, M. Kuhn, M. Lorenzi and R. Wattenhofer, "FromWeb to Map: Exploring the World of Music", *IEEE/WIC/ACM Int. Conf. on Web Intelligence*, 2008.

[14] O. Hilliges, P. Holzer, R. Klüber and A.Buts, "AudioRadar: A metaphorical visualization for the navigation of large music collections", in *Lecture Notes in Computer Science*, 2006.

[15] X. Serra, "Technologies to support the collaborative production of sounds: the example of Freesound.org", presented at the *Unlocking Audio 2* conference, British Library, London, UK, 16-17 Mar., 2009.

[16] W.-P. Wellinga and B. Planqué, "Xeno-canto: web-based sound collecting and identification", presented at the *Unlocking Audio 2* conference, British Library, London, UK, 16-17 Mar., 2009.

[17] M. Magas, M. Casey and C. Rhodes, "mHashup: Fast Visual Music Discovery via Locality Sensitive Hashing", *ACM SIGGRAPH*, 2008.

[18] M. Casey and M. Slaney, "Song Intersection by Approximate Nearest Neighbour Retrieval", *Proc. International Conference on Music Information Retrieval (ISMIR)*. Victoria (BC), 2006.

[19] C. Rhodes and M. Casey, "Algorithms for Determining and Labelling Approximate Hierarchical Self-Similarity", in *Proceedings of the International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[20] M. Casey, C. Rhodes, and M. Slaney, "Analysis of Minimum Distances in High Dimensional Musical Spaces", IEEE Transactions on Audio, Speech and Language Processing, 2008.

[21] P. Proutskova and M. Casey, "You call that singing? toward MIR tools for multi-cultural collections of ethnomusicological recordings," *Proceedings of the International Symposium on Music Information Retrieval*, 2009 (Under submission).

[22] J.-J. Aucouturier, F. Pachet, P. Roy, and A. Beurive, "Signal + context = better classification," *Proceedings of the International Symposium on Music Information Retrieval*, 2007.

[23] P. Proutskova, "Musical memory of the world - data infrastructure in ethnomusicological archives," *Proceedings of the International Symposium on Music Information Retrieval*, 2007.

[24] V. Grauer, "Echoes of our forgotten ancestors", in *The World Of Music*, vol. 2, 2006.

Towards the Disintermediation of Creative Music Search: Analysing Queries To Determine Important Facets.

Charlie Inskip
Dept of Information Science
City University London
c.inskip@city.ac.uk

Andy MacFarlane
Dept of Information Science
City University London

Pauline Rafferty
Dept of Information Studies
University of Aberystwyth

ABSTRACT

Creative professionals search for music to accompany moving images in films, advertising, television. Some larger music rights holders (record companies and music publishers) organise their catalogues to allow online searching. These digital libraries are organised by various subjective musical facets as well as by artist and title metadata. A facet analysis of a collection of written queries is discussed in relation to the organisation of the music in the bespoke search engines. Subjective facets such as Mood and Genre are found to be highly important in query formation. Unusually, detailed Music Structural aspects are also key. These findings are discussed in relation to disintermediation of this process. It is suggested that there are barriers to this, both in terms of classification and also commercial / legal factors.

I. INTRODUCTION

Music Owners, such as record companies and music publishers, attempt to exploit the recordings and compositions they control by encouraging their use in films, TV programs, commercials, websites and corporate presentations. The process of adding music to a moving image is known as synchronisation. Many Owners' collections are digitised and act as digital music libraries. Music Users such as ad agency creatives and music supervisors search for music for synchronisation. They generally deal direct with a number of expert intermediaries employed by the Owners, who interpret the query and perform searches of their own catalogues on the Users' behalf.

A number of Owners operate online search tools which are designed to disintermediate this process. In an investigation of the metadata presented by these bespoke Music Search Engines (MSEs) [1] to the User [2] it was shown that Bibliographic and Descriptive terms are used in their music classification schemes. Some of these such as Subject and Mood, are outside the traditional music classification paradigm.

Recent important studies in music user information need [3, 4, 5, 6] have focussed on consumers. This paper is part of ongoing research into a group of creative professionals who may have different information needs than recreational consumers. These professionals are choosing music on behalf of others, often to convey or reinforce a message. The search for music to accompany moving images is frequently an unknown item search. These professional Users often do not know specifically what they are looking for but they seem to have very clear ideas of what elements are important, such as Mood, Genre and Structure. In advertising they are also often looking for a suitable 30 second element, not the whole song.

The contribution of this paper is to analyse a selection of 27 written queries in order to inform systems development [6, 7]. In the next section the Methodology is presented. Section 3 discusses Findings, focusing on descriptive and bibliographic facets and additional ways of constructing and clarifying synchronisation queries. This is followed by a discussion on factors affecting whether the process may be disintermediated. Conclusions are summarised in Section 5.

II. METHODOLOGY

During the course of a wider research project looking at information needs and behaviour in the process of searching for music to accompany moving images [8, 9], 27 written queries ('*briefs'*) were collected. These briefs came from creative music searchers who are employed by advertisers or brands to find music to be used in advertising, on websites, in corporate presentations etc.. Five of these briefs related specifically to TV trailer use, 21 related to commercials (one was duplicated), one was for web-site use. Briefs are often sent by email to Music Owners when a music search is taking place and they attempt to encapsulate the search criteria for an unknown piece of music which will match existing footage. They are a rich source of information regarding the semantics of music search. They are often up to one page in length and include a range of detail explaining the Users search criteria to the catalogue Owner.

The aims of this approach were:

a) to investigate the semantics of creative music search;
b) to relate this to knowledge organization in existing bespoke music search engines;
c) to make observations on whether the process may be disintermediated.

The metadata used by a selection of MSEs had been examined previously [2]. The facets by which these systems are organised is divided according to whether it is bibliographic (eg Artist, Writer, Title) or descriptive (eg Mood, Subject, Genre). It was suggested that these systems are organized by taking a domain analytic approach – the Owners classifying the documents by purpose [10]. This required an analysis of User discourses to evaluate whether they match the way the MSEs are organized. In pursuit of substantiation of this proposal, and for the purposes of this paper, the 27

queries were analyzed in depth and examined for links with and differences to the MSEs organization.

The briefs were imported into NVivo 8 software [11]. This flexible Computer Assisted Qualitative Data Analysis (CAQDAS) package allows texts to be searched, coded, annotated and queried. Each brief was analysed word by word and phrase by phrase and coded according to facets derived from the previous MSE analysis (Artist, Title, Year, Genre, Subject, Mood etc). As new facets arose these were iteratively added to the coding list. When all of the briefs had been coded, the facets were ranked by frequency of appearances within the set of briefs (see Appendix).

The sections that had been coded were then examined in turn in order to consider the words that had been used within each code. For example, in the code 'Artist' there were 11 references in total. It is a simple process with NVivo to isolate these references and analyse their role within the discourse of the query. This approach enables the researcher to consider not just the words that are coded (eg 'Pink Floyd') but also the words *around* the coded words:

> *"i.e. a classic piece of music like Pink Floyd - We don't need no education - a track which is about rebellion."* (Brief 001)

The value of this discourse analytic approach [12] is that it is the words on either side of the coded words that help to explain the context of the reference. In the example above, if 'Pink Floyd' were the only words considered, we would not appreciate that this is partially a similarity request rather than a known item request. The User is not solely asking for Pink Floyd's 'We don't need no education' (sic – the correct song title is 'Another Brick In The Wall Pt 2'). Pink Floyd are partly being used as a similarity metric to give context to the request, although the User is also asking whether this song (or a version of it) is available.

A word frequency count was also performed on the entire query set. Irrelevant terms such as 'should', 'can', 'his', 'you' etc were discarded from the list. Facet terms and musical descriptors were concentrated on. Analysing this list gave deeper insight into the regularity of appearance of some key concepts. Again, the terms could be drilled down on using the software to reveal their context so they were not considered in isolation from the rest of the surrounding text.

III. FINDINGS

i. DESCRIPTIVE FACETS

By far the main facet used to describe the music being sought is that of Mood, which featured in 80% of the briefs. Positive descriptors such as *'charming', 'beautiful', 'fresh', 'playful', 'quirky', 'exciting'* far outweighed negative terms such as *'dark', 'uncertain', 'anxious', 'sinister'*. Notably these negative terms are mainly used as 'Exclude' terms in order to instruct the intermediary that music matching these would not be relevant for the search in question ("*Please do not pitch music with an overtly sinister, dark, or serious feel"*). Although a larger sample of queries could generate more 'negative' mood criteria it seems likely that as these queries are focussed on finding music for advertising, the users are looking for positive music moods to associate with and enhance consumers' opinions of their products. Mood has been used to match music to images since the time of silent movies [13] and advertising theorists are well aware of its value in selling products [14]. As a subjective facet it is not an ideal way to describe the elements of music that the user is looking for, as it is difficult to match the users meaning with that of the intermediary or, indeed, the system. However the use of Mood (this is specified either for the sought music or to describe the desired 'feel' of the finished advert) far outweighs any other term employed in this set of queries.

Unsurprisingly Genre (*"heavy metal", "rock", "pop", "lofi", "folky", "classical", "jazz"*) is mentioned in many queries as a guide to the music being sought. Although Genre definitions are fuzzy, generally they are agreed by a community and can be applied more successfully than more subjective terms such as Mood. Genre is a useful way for the search to be narrowed down, at least, and its widespread long term use in describing especially different forms of popular music as well as the three main genres (Art (Classical), Popular (Pop) and Folk) indicates it is an extremely valuable approach to music classification. The use of Genre terms can help as codes in describing music for a particular audience (products aimed at the youth market are often associated with contemporary pop), or which instruments would be appropriate (electric guitars do not often figure in classical).

Given the short length of the TV commercial, it is rare that a whole piece of music is used to accompany the footage, unless it is bespoke. The Users are looking for a specific element of a piece that they can use to convey their message to the viewer. They may discuss (in Music Structure) what musical elements they are looking for that may appear within a song: *"should have some quieter moments", "music evaporates into nothing", "build to a swelling, string-soaked chorus", "...with a crescendo in it"*. The word, *build,* in particular appears regularly in these queries and in other discourses surrounding these practices. These content-based criteria are very important to these Users on two levels. They help to convey the message to the viewer, and they also allow important Extra-Musical Factors (such as sound effects or voice overs) to be used successfully.

ii. BIBLIOGRAPHIC FACETS

While the use of subjective facets seemed to be key in communicating Users' music needs to Owners, a equal number of bibliographic facets are also employed. The benefit of factors such as Date/Period (of recording being sought), key words required in Lyrics, Tempo, Instruments featured in the track and Chart Position is that they are easily attached to music documents as metadata and can be more reliable search parameters.

The value of Date/Period is that it can be matched to target Audience demographics, as well as being used to refine a general description of a style of music. There are relatively frequent references to finding music that is "*contemporary*", while other briefs refer to decades rather than particular years:

> "*Please avoid 80s electronica, retro tracks, or anything that could be considered 'old skool'.*" (Brief 011)

> "*Instinctively we think that the track we need is probably from the 50's or 60's, maybe the 70's.*" (Brief 012)

Songs that include particular lyrics are discussed. Examples of these include:

> *"We are looking for lyrics which need to be relevant and carry the ad. Think along the lines of ideas / imagination / optimism / growth / design / drive / movement etc etc..."* (Brief 007)

> *"Lyrics and choruses involving sleep, eyes, waking, dreaming, touch, or some other direct link to the narrative, would be great."* (Brief 012)

However lyrics are not always important and often Instrumentals (no vocals) are requested. This use of instrumentals not only gives space to voice overs (VO) and sound effects (SFX) but recognises the creative nature of advertising and sophistication of the viewers who are well-versed in interpreting these short messages without relying on lyrical reinforcement.

Content-based musical facets such as tempo and instruments are occasionally mentioned in this sample of briefs. It is interesting to note that by far the most frequent tempo descriptor is '*upbeat*', a term indicating a positive mood as well as a faster than average tempo. This particular combination here of affective and structural facets into one descriptor is very effective shorthand which appears so frequently in interviews on the subject as to become a cliché (Inskip et al 2009). Users also mention key instruments (piano, guitar, strings, percussion) they wish to appear in the selected music.

Artist name is occasionally used, mainly as a similarity guide rather than a known item search:

> *"We are looking for a recognisable song that reflects a 'Happy Goodbye'. Think 'My Way' as performed by Frank Sinatra."* (Brief 023)

In fact it would not be easy for these MSEs to match items by similarity. They can only search catalogue they control and the example may not be within that control. Intellectual Property (IP) legislation can prohibit them from including material not under their ownership, restricting their ability to develop this type of search functionality.

Chart position, on the other hand, is easily available to Owners and is a simple way to measure 'familiarity'. If a User requests a familiar tune this means it is likely to have appeared in the sales charts so searching by chart position can be used in a 'familiarity' search.

Although they are often the most important factor in deciding whether a piece of music is used (or not), budgets are rarely revealed in queries:

> *"..budget can be stretched."* (Brief 001)
> *"..without being very costly!"* (Brief 017)
> *"Don't worry about budget please."* (Brief 024)

The expert interpretation of these Budget facets along with a knowledge of the brand's budgeting history and an understanding of which elements of the catalogue command different rates can lead to certain types of music being offered.

iii. VISUALS FACETS

Although most queries in this sample focussed on advertising, a small number were concerned with looking for music for websites or TV trailers. Mentioning the Format (ad, tv, website) in the query gives richer detail to the intermediary about the eventual use of the music being sought and is an additional clue to specific facets which may be of interest. These would include length (TV ads are normally 30 seconds long, while website uses may require the whole song, or a loopable section) and raise the issues of licensing – using a piece for a TV ad would require different permissions than web or TV trailer use. These may help the intermediary in narrowing down the search results to manageable levels. Other Visuals Facets, such as Project Title, Visuals Subject, Brand, Visuals Function and Visuals Available are also incorporated into the queries. These provide detailed contextual information for the intermediary and help to clarify the query further.

iv. QUERY CLARIFICATION

There are a number of phrases within the queries where the Users attempt to clarify their query by discussing the role of the music within the finished advert. Music Function appears frequently. The Users describe how they wish the music to interact with the visuals, or how they want the music to enhance their message:

> *"...juxtapose against this theme..."* (Brief 001);
> *"The music needs to complement the visuals without being too cold"* (Brief 003);
> *"...reflect the charm and playful nature of the spot"* (Brief 004);
> *"tidily juxtapose with the childlike imagery of the animatic"* (Brief 007)
> *"reflect the gliding motion of the journey"* (Brief 009).

The value in matching music to moving images is that one can enhance the other (in the case of advertising the music is designed to enhance the visuals, while with music videos it is the visuals that are designed to

enhance the music). It is not clear from the queries how this is evaluated, and while other research indicates this is often a 'gut feeling' decision based on experience and creativity, there is also a wealth of literature which discusses music and its use with film since the end of the 19th century from which users may draw. Clearly this type of criterion can only be evaluated once the music is played simultaneously with the image. 'Demo' versions of the final advert are frequently supplied along with the query in order to help the intermediaries in their search.

While the bulk of the text of the queries describes what the users are looking for, they often also clarify what would not suit. These Exclude elements again are designed to guide the search by narrowing down the results set:

> *"we want to avoid anything that makes it feel like a middle class day out at the shops"* (Brief 019);
> *"avoid anything too folky or dreamy", "something known will be tossed"* (Brief 025), *"it is important not to emotionalise things too much"* (Brief 026)

although careful interpretation by the intermediary is again required.

For the purposes of query clarification other intertextual references may be used, such as Films that use a particular type of music, Similarity to other suitable pieces or artists, and the option to the intermediary to offer a piece that does not appear to match the query but may be appropriate: this Left Field element reflects the subjective nature of this type of searching, and allows the expert intermediary to make a contribution to the search using their own experience and interpretation of the visuals:

> *"Please also feel free to suggest anything off brief that you think works well to picture, we are open to suggestion no matter how left field."* (Brief 007)
> *"But feel free to send other suggestions if you think they work."* (Brief 012)

There are many anecdotal examples of music being used in commercials that did not meet the original brief and were supplied by the intermediary as a 'Left Field' suggestion:

> *"She just threw that in as a kind of a random idea, and they went for it."* (Music Supervisor)
> *"Sometimes you have to come up with something that's completely off the wall that the director won't have thought about"* (Music Publisher)

IV. WHITHER DISINTERMEDIATION?

We have seen that music briefs describing music which will accompany moving images incorporate a range of content-based music facets and additional contextual detail. Some of these can be matched with bibliographic metadata, while others (tempo, instruments, and especially structural facets such as crescendo or build, can be retrieved using state-of-the-art content-based retrieval. However a large number of subjective facets are used, relying on a shared understanding of the meaning between the User and Owner. The expert intermediaries employed by the Owners are well-versed in interpreting these queries, may have the opportunity to discuss them in detail with the Users and sometimes view a copy of the piece of film in question to help in their search.

Building a searchable digital library that suits this verbose and subjective type of request is not an easy task. Granted, some of the bibliographic facets can be dealt with by applying suitable metadata fields, but these have to be accurate if they are to be of any value. Songs are often classified by inexperienced humans leading to inconsistent and unreliable metadata [2]. With regard to the bibliographic facets this is a solvable problem. However it appears from our analysis that these searches rely more on descriptive metadata and detailed musical structural facets than factual bibliographic detail. This means that if the process is to be satisfactorily disintermediated the focus needs to be on successfully matching query terms with relevant and suitable metadata combined with successful feature extraction and state-of-the-art content-based retrieval techniques.

i. MOOD

Let us consider Mood. Users employ a wide range of words:

> *'charming', 'beautiful', 'fresh', 'playful', 'quirky', 'exciting', 'dark', 'uncertain', 'anxious', 'sinister'.*

The MSEs also use a wide range of Mood descriptors from controlled vocabularies, which are presented to the User for them to select the most appropriate to their search:

Figure 1 MSE moods (Inskip et al 2009)

Encouraging a User to select a Mood from a controlled vocabulary, rather than asking them to input it as a

keyword means it is the Owner who is also involved in encoding the query to match the system, rather than the User encoding the query to match the music being sought. This can remove the creative element of the search from the User and may dissuade them from attempting to perform their search online. Clearly, if the Mood choices are to be presented to Users then it is important to investigate how they determine whether a piece is '*charming'*, '*beautiful'* or '*fresh'* and developing the controlled vocabulary accordingly. This applies equally to Genre, although as previously stated, Genre is less subjective than Mood. The variation in interpretations of the cultural meanings of music reinforces the value of taking a domain analytic approach when developing music search tools [15].

ii. MUSIC STRUCTURE

It is highly unusual that these searches are focusing on a small element of lengthier works. The viewer only hears up to 30 seconds of music in a TV commercial, probably less. Most popular songs last around 3 minutes. It is very important in the search process that the part of the piece that matches the search criteria is found. Songs may vary in mood, tempo, may have crescendos at the middle or the end. The whole song has to be listened to in order to find out whether it includes a relevant section. Advertising creatives have little time and would benefit from being presented with the 30 second element of a song that matches the query rather than be forced to listen to an entire song. When the human intermediary does the search s/he may know where in a piece the '*build'* takes place and possibly will direct the User to this section of the music. Disintermediation could perform a similar function. Content-based retrieval tools that search music signals for Music Structural facets such as crescendos, solos, and specific instrument onsets would be of particular value in this area.

iii. COPYRIGHT AND COMPETITION

It may be that a 'global' search system may help the searchers who are time poor by saving them doing the same search on multiple services. However there are legal and business hurdles to this solution. Record companies may not use lyrics without permission – they are controlled by music publishers. Conversely, music publishers may not own the recordings of the compositions they control and would need permission to include them in their online catalogues. If we combine this problem with the fact that these services are designed for the exploitation of catalogue in a highly competitive industry then collaboration between Owners is difficult. However it is not unsurpassable. There is currently one service (www.ricall.com) which unifies a selection of client catalogues in an attempt to simplify the problematic element of the search process relating to Users having to use numerous interfaces for one search.

V. CONCLUSION

Anecdotal evidence suggests that, historically, although there is the will to disintermediate the process, there have not always been the resources or the technology. A number of MSEs were developed up to five years ago in a rush to compete for business. They have not all been updated to keep up with web and search engine technology, although there are exceptions with some services currently in re-development and not available for analysis.

It appears that although the Music Owners are designing search tools for Users who wish to search online [2] the possible mismatch between the Users' approach and that of the Owners must be considered. If successful disintermediation of these services is to take place then the Users and their contexts have to be considered in detail.

It is hoped that the long term value of this research is not only to the commercial users of creative search represented in this study, but to wider users in the general public who wish to choose music to accompany home videos, slide shows and presentations. Disintermediation of the search process is a reality and is a central aim of the retrieval community. We are interested in whether currently available MSEs, only available for commercial use, match real information needs of a small researchable group of Users. Our results are regularly circulated amongst the participants and also in academic and professional music information retrieval and library and information management circles in the hope that they benefit not just the commercial world but inform wider systems development.

ACKNOWLEDGEMENTS

We would like to thank all the participants in our research for providing us with briefs and being so free with their valuable time. Charlie Inskip gratefully acknowledges financial support from AHRC for this PhD research.

REFERENCES

[1] Nanopoulos, A., Rafailidis, D., Ruxanda, M, & Manolopoulos, Y. (2009) Music search engines: Specifications and challenges, Information Processing and Management 45(3) pp 392-396

[2] Inskip, C., Macfarlane, A. & Rafferty, P. (2009) Organizing Music for Movies. Proceedings of International Society for Knowledge Organization (UK) Content Architecture conference, London, UK, 22-23 Jun 2009

[3] Bainbridge, D., Cunningham, S., and Downie, J. (2003) How people describe their music information needs: a grounded theory analysis of music queries. Proceedings of the 4th International Conference on Music Information Retrieval, Oct 26-30, Baltimore, Maryland.

[4] Cunningham, S., Bainbridge, D. and McKay, D. (2007). Finding new music: a diary study of everyday encounters with novel songs. Proceedings of the 8th International Conference on Music Information Retrieval, Sep 23 – 27, Vienna, Austria

[5] Kim, J. & Belkin, N. (2002), Categories of Music Description and Search Terms and Phrases Used by Non-Music Experts, Proceedings of the 3rd International Conference on Music

Information Retrieval, Paris, France, Oct 13-17 2002.

[6] Lee, J., Downie, J.S.; Jones, M.C. (2007) Preliminary Analyses of Information Features Provided by Users for Identifying Music, Proceedings of the 8th International Conference on Music Information Retrieval, Vienna, Austria, Sep 23-27 2007.

[7] Downie, S. and Cunningham, S. (2002) Toward a theory of music information retrieval queries: system design implications. Proceedings of the 3rd International Conference on Music Information Retrieval, Oct 13 – 17, Paris, France.

[8] Inskip, C., Macfarlane, A. & Rafferty, P. (2008a) Meaning, communication, music: towards a revised communication model. Journal of Documentation, 64(5) pp 687-706

[9] Inskip, C., Macfarlane, A. & Rafferty, P. (2008b) Music, Movies and Meaning: Communication in Film-Makers Search for Pre-Existing Music, and the Implications for Music Information Retrieval. Proceedings of the 9th International Conference on Music Information Retrieval, Philadelphia, PA, Sep 14-18 2008

[10] Hjørland, B. and Nissen Pedersen, K. (2005) A substantive theory of classification for information retrieval. Journal of Documentation 61(5) pp582-597.

[11] QSR (2009) NVIVO Product Homepage, <http://www.qsrinternational.com/products_nvivo.aspx> last accessed 2 Sep 09

[12] Paltridge, B. (2006) Discourse Analysis. Continuum, London and New York.

[13] Rapee, E. (1924) Motion Picture Moods for Pianists and Organists. Arno Press, New York, 1974 reprint.

[14] Alpert, J. & Alpert, M. (1990) Music Influences on Mood and Purchase Intentions, Psychology and Marketing 7(2) pp 109-133

[15] Abrahamson, K.T. (2003) Indexing of Musical Genres: An Epistemological Perspective Knowledge Organization 3 / 4 pp 144-169

Brief 001 – 027 (2009) Private documents supplied by anonymous creative music searchers

APPENDIX

Music Facets	**References**	**Type**
Mood	130	Subjective
Genre	39	Subjective
Music Structure	21	Subjective
Date / Period	20	Objective
Audience	16	Objective
Lyrics	14	Objective
Artist	11	Objective
Tempo	10	Objective
Instrument	9	Objective
Extra-musical	9	Objective
Song Title	7	Objective
Chart position	5	Objective
Budget	4	Objective
Version	3	Objective
Music Style	3	Subjective
Length	3	Objective
Instrumental	3	Objective
Clearable	3	Objective
Vocal	1	Objective
Territory	1	Objective
Song Subject	1	Subjective
Other intertextual refs	1	Subjective
Exploitation	1	Subjective

Visuals Facets	**References**	**Type**
Format (ad, film, tv)	25	Objective
Project Title	18	Objective
Visuals Subject	16	Objective
Brand	13	Objective
Visuals Function	6	Objective
Visuals Available	6	Objective

Query clarification	**References**	**Type**
Music Function	47	Subjective
Exclude	22	Subjective
Film title	9	Objective
Similarity	6	Subjective
Left Field	4	Subjective

Preserving today for tomorrow: A case study of an archive of Interactive Music Installations

Federica Bressan
Department of Computer Science
University of Verona
Strada le Grazie 15
37134 Verona
federica.bressan_01@univr.it

Sergio Canazza, Antonio Rodà
Lab. AVIRES
University of Udine
Via Margreth 3
33100 Udine
{sergio.canazza, antonio.roda}@uniud.it

Nicola Orio
Information Management
Systems Research Group
University of Padova
Via Gradenigo 6a, 35100 Padova
orio@dei.unipd.it

***Abstract*—This work presents the problems addressed and the first results obtained by a project aimed at the preservation of Interactive Music Installations (IMI). Preservation requires that besides all the necessary components for the (re)production of a performance, also the knowledge about these components is kept, so that the original process can be repeated at any given time. This work proposes a multilevel approach for the preservation of IMI. As case studies, the *Pinocchio Square* (installed in EXPO 2002) and the *Il Caos delle Sfere* are considered.**

I. Introduction

Since the 1970s, computers have become the main tools to explore the acoustic world by deconstructing the sound matter at the level of elementary attributes of signals relying on different analysis techniques, while programming languages and synthesis techniques have allowed a nearly total control over the sonic material.

During the 1990s physical and spectral modeling was enhanced and various interfaces for musical creativity were developed and widely spread. The exploration of different levels of complexity, from physics to phenomenology and semantics, was made possible by improved tools that balanced cost and efficiency.

The musical works created with the above-mentioned technologies are represented in a *fixed* form (i.e. recorded music). In this sense, the preservation problem has reference to the preservation of the audio documents (i.e. audio signal, metadata and contextual information[1]).

[1]In this context, as it is common practice in the audio processing community, we use the term metadata to indicate content-dependent information that can be automatically extracted by the audio signal; we indicate as contextual information the additional content-independent information.

More recently, the role played by multimediality within the performing arts has become more and more important, most notably in the music field, where the opening to interaction between images and sounds fostered the experimentation of new expressive solutions by the artists. Along with the evolution of the creative processes, the development of increasingly more sophisticated audiovisual technologies affected the ways in which the public acknowledges the artistic event, its habits and aesthetics. With innovative solutions for mutual interaction, music, dancing and video turned into "expanded disciplines", thus giving birth to "Interactive Music Installations" (IMI). At present, the spotlight is on physically-based sounding objects that can be manipulated and embodied into tangible artifacts that support continuous interaction. Technology offers the tools to detect and control sets of variables which modify specific elements within a system (sensors), it relies on an enhanced computational power and it pays special attention to design strategies.

Generally IMIs are able to detect and analyze motion, speech, sounds produced by one or more users, in order to control sound synthesis, music and visual media (laser effects, video, virtual avatar) in real-time, and to modify the environment (setting and mobile robots). IMIs observe and act within a certain environment, modifying their own structure and responses in a dynamic way according to the users' behavior (same gestures have different effects in different settings). Basically IMIs are able to retrieve general features in a Gestalt-like approach to the detection of sound and motion.

Unlike virtual reality and multimediality, IMIs do not aim at deceiving the human sensory system, but to extend reality, allowing the user to interact with an expanded

world by means of technology, mainly for artistic and esthetic purposes. Particularly important in this sense is the SAME project (http://www.sameproject.eu/) that aims at creating a new end-to-end networked platform for active, experience-centric, and context-aware active music listening/making.

Besides providing scientific and technological requirements for the creation of new forms of musical dramaturgy, the playback and recording devices for audio/visual reproduction is the one to suit best the preservation and the spreading of the moment of the performance. Collecting several recordings of an art work, shot under different circumstances, allows a thorough documentation of the history of the work itself. Furthermore, recordings are the only means to keep memory and to increase the value of those works which were performed only once.

Digital tools makes it possible to (i) watch over and over again an event after it was performed and to (ii) reorganize its elements with hyper-textual and interactive processes, for example by enriching the video-recording with interviews and backstage footage. Besides, shooting an event with several cameras introduces a multiple point of view which is impossible to have in a regular theater set, where the experience is usually static and tied to the features of the venue. Such possibilities clearly point out that recordings are artificial reconstructions that by no means can take the place of the original performance.

In the field of audio documents preservation some relevant guidelines have been sketches along the years, but most questions regarding the safeguard and the preservation of IMIs remain unanswered, as the (European) regulations in force do not provide for specific care or legislative obligations. One of the few work related to the IMI preservation is the ontology approach used in [1], [2] to describe an IMI and its internal relationships to support the preservation process. The proposed ontology is an extension of the CIDOC Conceptual Reference Model (CIDOC-CRM), which is an ISO standard for describing cultural heritage [3]–[5].

The study of several projects related to the preservation of virtual artworks gave rise to some interesting considerations: the Database of Virtual Art in Germany [6] is aimed at archiving expanded documentation about virtual artworks; the Variable Media Network [7] encourages artists, in the field of conceptual, minimalist and video art, to define their artworks medium-independently so that the artworks can be recreated at another time, when the medium becomes obsolete; the Variable Media Network also aims at developing the tools, methods and standards required to implement this strategy. Altogether, research in preservation of artworks has investigated different preservation methods. However, there has been no definitive answer on which one is the best approach for preservation of artworks in general and IMI in particular. In our field, the preservation of various recordings (such as still images, audio and video) remains necessary, but not sufficient. Motions of performers as well as the setting of performance space also need to be captured and preserved. Our differentiation is that we strongly believe that not only an exact experience of a performance must be recreated, but also the interactions must be archived for future analyses, studies and fruition.

In this sense, this work presents the problems addressed and the first results obtained by one of the first projects completely dedicated to the preservation of Interactive Music Installations. After an overview of the problems related to the IMI preservation (Sec. II), Sec. III, considering [2], presents a multilevel approach for the creation of an IMI archive. Finally, it describes some experimental results obtained in the preservation of two installations, the first produced for EXPO 2002 and the second presented in a number of Italian venues (Sec. IV).

II. Interactive Music Installation: preserving the *experience*

The meaning of an IMI inextricably lies in the interaction among people, objects, and environment [8], in this sense IMIs must have (at least) three characteristics: (i) a phenomenological essence, that is meanings and structures emerge as experienced phenomena; (ii) an embodiment; (iii) the human interaction .

As mentioned above, music works that were written or recorded in a fixed form mainly refer to problems that lie within the scope of the audio documents preservation. IMIs demand a different challenge, that is to *preserve the interaction*. This involves the preservation of: audio signal, metadata, contextual information and *experience*.

In the IMI context, audio (non-speech) signal contributes to the multimodal/multisensory interaction, which can have several meanings. It represents the main goal of the IMI in itself when it is the vehicle for the music performance, but the audio channel can also be used to communicate events and processes, to provide the user with information through sonification, or to give auditory warning. This communication is related to the physical world where the IMI is immersed. Depending on the environment where the IMI is set, the same sound can elicit different levels of representation. In visual

perception, space can be considered an indispensable attribute [9], i.e. a prerequisite for perceptual numerosity. The Time-Frequency plane (with spectra, envelopes, etc.) is the principal dimension of auditory perception, where auditory spatial perception is in the service of visual perception: *Ears lead Eyes*.

All this considered, it should be now clear that it is not possible to use the protocols and instruments defined in the field of multimedia documents preservation: in fact what we preserve are not (only) audio and video documents. So what do we preserve? *Useful information* [10] is the answer to the question. In this sense, we must preserve at least two sound attributes: source localization and the sensation of *ambience*. This is not necessarily realism, but rather *Fidelity in Interaction*.

III. Preserving (faithfully) the interaction: a multilevel approach

As expressed in [11], "preservation is the sum total of the steps necessary to ensure the permanent accessibility – forever – of documentary heritage". In most forms of traditional artistic expressions, the artifact coincides with the work of art. In Interactive Music Installations, however, it is less clear where the art lies, as it usually "happens" during the process of interaction. Preserving the installation as it was presented to the public falls within the scope of current approaches to preservation, yet the creative process is something impossible to freeze. Therefore, in this sense, the preservation is an ongoing process: nothing has ever been preserved: it is being preserved. In our opinion, in order to grant a permanent access, which is the ultimate objective of preservation, a combination of approaches is desirable. The *model* that underlies the interaction process is essential. This leads to different levels of preservation, with different purposes and ways to be performed. Moreover, we must consider that two forms of preservation exist: static where records are created once and not altered and dynamic where records keep changing. The next paragraphs describe a multilevel preservation approach, which is summarized in Tab. I.

Preserve the bits – Each part of the original installation that can be directly preserved. All the data are kept in the original format (the problem of their interpretation is a matter aside), and the risk of introducing alterations must be avoided. It is a physic and static form of preservation, and it works in the original contexts. It still requires metadata preservation and refreshing. There can be physical and digital objects. Preserving the digital objects will follow the general conceptual model defined in the OAIS Reference Model [12]. This means that each digital object will be accompanied by its representation information, preservation description information (i.e. reference, context, provenance and fixity) and descriptive information. The *performance* objects also need another layer of information about itself as a whole, including information about: i) various activities necessary for the performance and their temporal relationships; ii) necessary components of the performance and how these components can possibly be linked together to create the performance. The most challenging problem is to preserve the knowledge about the logic and temporal relationships among individual components so that they can be properly assembled into a performance during the reconstruction process.

Preserve the data – Technical notes, comments and useful information about the realization of the installation. Includes a high level descriptions of algorithms and models. No special attention is paid to presentation and context. It is a physic and dynamic form of preservation, as it could be necessary to use new design languages, possibly developed on purpose.

Preserve the record – Any element that was modified or updated in respect of original installation. Includes reinterpretation of the patches and information about the context. Costs are balanced against utility, and appearance is not necessarily preserved. It is a logical and static form of preservation. Some risks are tolerated, if not necessary (e.g. migration to new informatics environments). The results of each intervention should be checked using philological instruments. This level needs another layer of information about activities *actually* performed in the performance (the actual activities performed during a performance can be different from the necessary activities mentioned at the *bit* level).

Preserve the experience – Any document that bears witness to some aspect of the installation. Includes hardware, software, interviews, audio/video recordings, usability tests of the original system, as well as information about people (composers, directors, performers, technicians) involved in the original performance and their roles. Resembles a museum-like approach, that aims at keeping track of the history of the installation. It may require emulators or old computers (emulators reproduce a whole system, or a program; system emulators require us to use old interfaces; program emulators may use current interfaces) and/or migration in new systems (*reinterpretation*). Although emulation has been known as a method that could create the original look and feel of the work [13], [7] showed that this was not easily

achievable, owing to many differences between the original hardware platforms and their emulated counterparts, such as CPU speeds, as well as looks and feels of the new hardware platforms. Reinterpretation seems to be a preferred approach, where the artworks could be entirely encoded using a symbolic notation [14]. It is a logical and dynamic form of preservation, necessary for a long term action ($>$ 30 years). Its existence may be more important than the content, although it does not help reuse, it is expensive and not oriented to fruition.

The rapid obsolescence of the new technologies employed in Interactive Music Installations complicates the maintenance of hardware/software (bits and data). Besides, some installations base their interaction on objects that are meant to be used or consumed by the users (food, perishable material, etc.). This implies that parts of the installation may literally not exist anymore after the interactive process is over (no bits to preserve). The approaches described above do not require sequentiality, as they basically pursue different goals. Quite the contrary, they may occasionally show overlapping contents.

IV. Cases study

A. Pinocchio Square in EXPO 2002

The multimedia exhibition *Pinocchio Square* has been presented at the EXPO 2002, at Neuchatel (CH) in the period May-October 2002. Scientific partners were Sergio Canazza and Antonio Rodà, at that time, members of Computational Sonology Centre of the University of Padova, while involved artists were Carlo De Pirro (1956-2008: music composer, professor at the Conservatory of Music "Venezze" in Rovigo, from 1982 to 2008) and Roberto Masiero (artistic director, Professor in History of Architecture at the University of Venice). This IMI was motivated also by scientific aims, which were to test some movement analysis patches, for the measurement of high level features and to investigate some kinds of mapping between movements and music and its effect on a children audience. Moreover, an additional goal was to test the reliability of the Eyesweb environment (www.infomus.org/EywMain.html) during a long time performance (18 hours/day; 6 months).

1) Concept: In the contest of EXPO 2002, Switzerland built some wood platforms in its four major lakes. Each platform (80 x 30 meters) was dedicated to an artistic topic or to a scientific discipline. Our exhibition was installed in the platform dedicated to Artificial Intelligent and Robotics. The system was made by a room for children, like a magic room, in which each gesture becames sound, image, color. In this system the visitors are been involved in a communication of expressive and emotional content in non-verbal interaction by multisensory interfaces in a shared and interactive mixed reality environment. From the gestures of the visitors (captured by two video-cameras) are extracted the expressive content conveyed through full body movement. Mapping strategies will convey the expressive content onto a multimedia output (audio and video). The system focused on full-body movements as primary conveyors of expressive and emotional content. This approach implies a new and original consideration of the role that the physical body plays with respect to interaction. During the 6 months, the exhibition was visited by a large audience: the Swiss organization reckons about 4 millions of visitors. The exhibition was very appreciated in its artistic content.

2) Technical description: The setup was based on two video-cameras to capture full-body movements inside the room and two video-projectors to render the real-time video-processing in the room. Computation was carried out with a computer cluster (using Pentium III 800MHz, 128 MB Ram, HDD 16GB, Windows 2000) composed by: two PC to process the video captured information, called V1 and V2, and one PC to render audio content, called A. Sound was amplified by five loud-speakers (4 independent channel + 1 subwoofer). All the three PCs has installed and running Eyesweb 2.4.1 for both audio and video processing.

Description of the employed patches.

Patch 1. This patch is installed on PC V1 and is dedicated to the processing of the video, captured by the first video-camera. The patch is divided in two parts: the first, analyzes the video streaming, in order to calculate some high level features; the second, implements several algorithm for the real time processing of the video-streaming . The patch is connected via MIDI to PCs V2 and A.

3) Preservation Process: During the first 2 months of the preservation project carried out in CSC (University of Padova), the *bits preservation* task has been carried out: we created the conservative copies of all the documents related to the installation (software, textual, audio and video documents, etc.). In addition we described algorithms and models used by the installation (*preserving the data*). Finally, we migrated the patches in new Eyesweb environment in order to run the installation using new hardware with the Windows XP and Vista O.S. (preserving the record): figure 1 shows a frame of a running patch. We are carrying out the *experience preservation* step: we collected the original hardware and

Preservation approach	Static/Dynamic	Physic/Logic	Life expectancy (Years)
Preserve the Bits	Static	Physic	$5-10$
Preserve the Data	Dynamic	Physic	>30
Preserve the Record	Static	Logic	$10-20$
Preserve the Experience	Dynamic	Logic	>30

TABLE I
CHARACTERISTICS OF THE MULTILEVEL PRESERVATION APPROACH IN THE IMI FIELD

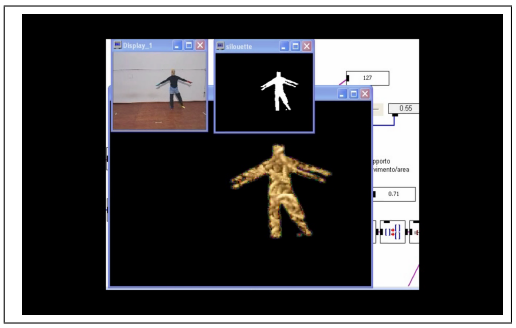


Fig. 1. The *vermi* (worms) patch migrated in new environment, with new release of Eyesweb and Windows XP.

software, and some interviews to the authors. The more difficult task is probably the re-creation of the original space: the magic room (with the same material). We have the original plan: up to now, lack of an adapted museum space and financial funds.

B. Il Caos delle Sfere: Be a Pianist with 500 Italian Lire

The interactive music installation *Il Caos delle Sfere: Become a Pianist with 500 Italian Lire* has been presented for the first time at the "Biennal of the Young Artists of Europe and Mediterraneo" Giovani Artisti di Europa e del Mediterraneo" in Rome, 1999; afterwards the exhibition toured in other artistic manifestations until year 2004. Scientific and technical partners were Nicola Orio and Paolo Cogo, at that time, members of Computational Sonology Centre of the University of Padova, while Carlo De Pirro was the involved artist. Although this IMI did not have scientific aims, it has been based on the results of a joint research work on music interaction called "Controlled Refractions" [15], on the interaction between a pianist and a computer through a music performance.

1) Concept: The basic idea was to use a common gaming machine, such as a electronic pinball, to control an automatic performance played on a Disklavier. The kind of interaction introduces a large amount of unpredictability on the obtained sound, because normal users have only a loose control on the "silver ball". Given that normally all the electronic pinballs give auditory feedback to the user, the basic idea of the composer was to avoid a one-to-one mapping between the objects hit by the ball and the generated sound. Moreover, the composer decided that a good player should be rewarded with a more interesting and complex performance than a naïve player. To this end, the amount of interaction varied according to the evolution of the game. The more levels were reached, the more complex and virtuosistic was the generated performance on the Disklavier. The game starts with some pre-written sequences; when the user changes to a new level, some automatically generated sequences start to play while the user partially controlled depending on the kind of targets he is hitting. At new level the style of automatic sequences changes, so it does the way the user can control them.

2) Technical Description: The installation was based on a popular electronic pinball "The Creature from the Black Lagoon". The choice of this particular pinball, which is one of the first to introduce the concept of different levels of a match where the player has to achive a number of goals that correspond to levels in the game, partially influenced the technical and artistic choice as explained in the previous section. The first technical goal was to monitor the game (levels and targets) minimizing the need to interface with the preexisting electronic. To this end, it has been chosen to split the signal coming from the pinball switches to track the targets hit by the ball. The level of the game was monitored by splitting the signal going to the pinball lights. It can be noted that the in this way it is only possible to estimate the level of the game and that some of the features (i.e. the actual number of points) have been neglected. The acquisition was made through an electronic circuit that has been designed ad hoc, which sends to the PC the information about switches and lights through the parallel port. A PC running Windows acquires the data through the parallel port, processes the information through a software developed in C by the technical team. The software is able to play, generate, and modify melodic sequences according to the indication provided by the composer. The result is sent to the Disklavier through the

MIDI port, using the environment Midishare developed by Grame (http://www.grame.fr/) and now available open source. The Disklavier, which could be either a grand or a upright piano depending on the place of the installation, was the only sound source. The goal was to minimize the presence of digital media, for instance, the PC has almost no graphical user interface.

3) Preservation Process: The *bits preservation* task has been partially carried out by creating conservative copies of all the software, the music sequences to be played at different levels. The electric scheme of the acquisition board has been digitized as well, while the notes of the composer about the installation are still in the process of being gathered. As regards the *data preservation*, a small team has been created for the definition of the used algorithms that, being developed in direct collaboration with the composer with a try-and-error approach, have been left mostly undocumented. Due to technical choices made about ten years ago, in particular the use of the parallel port and of a proprietary compiler which is no more supported, the *record preservation* is still an open issue. The *experience preservation* has to deal with the fact this IMI was not conceived for a particular space, and in fact it has been presented at different occasions from small indoor to large outdoor places. On the other hand, the experience preservation is directly related to the correct functioning of the electronic pinball, which has been completely restored.

V. Conclusion

Museums are experienced in describing and preserving things and events from the past. As for IMIs, out goal should be to maintain usable documents and events for the future. This work proposes a multilevel approach for the creation of an IMI archive. Some experimental results obtained in the preservation of the Installations *Pinocchio Square* and *Il Caos delle Sfere* created by Carlo De Pirro are described.

Acknowledgment

This work was partially supported by the Fondazione Arena di Verona and the University of Verona (Italy) within the framework of the joint cooperation project REVIVAL, 2009-2010, and by the University of Padova within the project "Analysis, design, and development of novel methodologies for the study and the dissemination of music works".

This paper is affectionately dedicated to the memory of the composer Carlo De Pirro (1956-2008), a leading researcher in Computer Music and an outstanding teacher whose brightness and kindness we will always remember.

References

[1] K. N. ng, T. Vu Pham, B. Ong, A. Mikroyannidis, and D. Giaretta, "Preservation of interactive multimedia performances," *International Journal of Metadata, Semantics and Ontologies*, vol. 3, no. 3, pp. 183–196, 2008.

[2] ——, *Metadata and Semantics*. Springer US, 2009, ch. Ontology for Preservation of Interactive Multimedia Performances.

[3] M. Doerr, "The cidoc crm – an ontological approach to semantic interoperability of metadata," *AI Magazine*, vol. 24, no. 3, pp. 75–92, July 2003.

[4] ——, "Increasing the power of semantic interoperability for the european library," *ERCIM News*, vol. 26, pp. 26–27, July 2006.

[5] T. Gill, "Building semantic bridges between museums, libraries and archives: the cidoc conceptual reference model," *First Monday*, vol. 9, no. 5, 2004. [Online]. Available: http://firstmonday.org/issues/issue9_5/gill/index.html

[6] H. Grau, "The database of virtual art: for an expanded concept of documentation," in *International Cultural Heritage Informatics Meeting: Proceedings from ICHIM03*. Paris, France: Ecole du Louvre, 2003. [Online]. Available: http://www.archimuse.com/publishing/ichim03/016C.pdf

[7] A. Depocas, J. Ippolito, and C. Jones, Eds., *Permanence Through Change: The Variable Media Approach*. Guggenheim Museum, 2003. [Online]. Available: http://www.variablemedia.net/pdf/Permanence.pdf

[8] P. Dourish, *Where The Action Is: The Foundations of Embodied Interaction*. MIT Press, 2001.

[9] M. Kubovy and D. V. Valkenburg, "Auditory and visual objects," *Cognition*, vol. 80, no. 1-2, pp. 97–126, 2001.

[10] S. Barrass, "Auditory information design," Ph.D. dissertation, The Australian National University, 1998.

[11] R. Edmonson, *Memory of the World: general guidelines to safeguard documentary heritage*. UNESCO, 2002.

[12] *Reference Model for an Open Archival Information System (OAIS)*. Consultative Committee for Space Data Systems, 2002. [Online]. Available: http://public.ccsds.org/publications/archive/650x0b1.pdf

[13] J. Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. Council on Library and Information Resources*. Council on Library and Information Resources, 1998. [Online]. Available: http://www.clir.org/pubs/reports/rothenberg/contents.html

[14] R. Rinehart, "A system of formal notation for scoring works of digital and variable media art," in *Annual Meeting of the American Institute for Conservation of Historic and Artistic WorkMustica Technical Documentation Server for Ircam Musical Works*, 2004.

[15] N. Orio and C. De Pirro, "Controlled refractions: A two-level coding of musical gestures for interactive live performance," in *Proceedings of the International Computer Music Conference*. Ann Arbor, MI: ICMC, 1998, pp. 88–91.

Multi-modal Analysis of Music: A large-scale Evaluation

Rudolf Mayer
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria
mayer@ifs.tuwien.ac.at

Robert Neumayer
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
neumayer@idi.ntnu.no

***Abstract*—Multimedia data by definition comprises several different types of content modalities. Music specifically inherits e.g. *audio* at its core, *text* in the form of lyrics, *images* by means of album covers, or *video* in the form of music videos. Yet, in many Music Information Retrieval applications, only the audio content is utilised. Recent studies have shown the usefulness of incorporating other modalities; in most of them, textual information in the form of song lyrics or artist biographies, were employed. Following this direction, the contribution of this paper is a large-scale evaluation of the combination of audio and text (lyrics) features for genre classification, on a database comprising over 20,000 songs. We present the audio and lyrics features employed, and provide an in-depth discussion of the experimental results.**

I. Introduction

With the ever-growing spread of music available in digital formats – be it in online music stores or on consumers' computer or mobile music players – Music Information Retrieval (MIR) as a research area dealing with ways to organise, structure and retrieve such music, is of increasing importance. Many of its typical tasks such as genre classification or similarity retrieval / recommendation often rely on only one of the many modalities of music, namely the audio content itself. However, music comprises many more different modalities. Text is present in the form of song lyrics, as well as artist biographies or album reviews, etc. Many artists and publishers put emphasis on carefully designing an album cover to transmit a message coherent with the music it represents. Similar arguments also hold true for music videos.

Recent research has to some extent acknowledged the multi-modality of music, with most research studies focusing on lyrics for e.g. emotion, mood or topic detection. In this paper, we apply our previous work on extracting rhyme and style features from song lyrics, with the goal of improving genre classification. Our main contribution is a large-scale evaluation on a database comprising over 20.000 songs from various different genres. Our goal in this paper is to show the applicability of our techniques to, and the potential of lyrics-based features on a larger test collection.

The remainder of this paper is structured as follows. In Section II, we briefly review related work in the field of multi-modal music information retrieval. Section III will outline the audio and lyrics features employed in our study. In Section IV, we describe our test collection and its most interesting properties, while in Section V we discuss the results in genre classification on this collection. Finally, Section VI will give conclusions and an outlook on future work.

II. Related Work

Music Information Retrieval is a sub-area of information retrieval, concerned with adequately organising, structuring and accessing (digital) audio. Important research directions include for example similarity retrieval, musical genre classification, or music analysis and knowledge representation. Comprehensive overviews of the research field are given in [1], [2].

The still dominant method of processing audio files in music information retrieval is by analysis of the audio signal, which is computed from plain wave files or via a preceding decoding step from other wide-spread audio formats such as MP3 or the (lossless) Flac format. A wealth of different descriptive features for the abstract representation of audio content have been presented. Early overviews on content-based music information retrieval and experiments is given in [3] and [4], focussing mainly on automatic genre classification of music. In this work, we employ mainly the Rhythm Patterns, Rhythm Histograms and Statistical Spectrum Descriptors [5], which we will discuss in more detail in Section III. Other feature sets include e. g. MPEG-7 audio descriptors or MARSYAS.

Several research teams have further begun working on adding textual information to the retrieval process, predominantly in the form of song lyrics and an abstract vector representation of the term information contained in text documents. A semantic and structural analysis of song lyrics is conducted in [6]. The definition of artist similarity via song lyrics is given in [7]. It is pointed out that acoustic is superior to textual similarity, yet that a combination of both approaches might lead to better results. Another area were lyrics have been employed is the field of emotion detection and classification, for example [8], which aims at disambiguating music emotion with lyrics and social context features. More recent work combined both audio and lyrics-based feature for mood classification [9]. Other cultural data is included in the retrieval process e.g. in the form of textual artist or album reviews [10].

A multi-modal approach to query music, text, and images with a special focus on album covers is presented in [11]. First results for genre classification using the rhyme features used later in this paper are reported in [12]; these results particularly showed that simple lyrics features may well be worthwhile. This approach has further been extended on two bigger test collections, and to combining and comparing the lyrics features with audio features in [13].

III. Audio and Lyrics Features

In this section we describe the audio and lyrics features employed in our experiments. The former comprise Rhythm Patterns, Statistical Spectrum Descriptors, and Rhythm Histograms. The lyrics features are bag-of-words features computed from the terms occurring in the songs, features describing the rhyming structure, features considering the distribution of certain parts-of-speech, and text statistics features.

A. Rhythm Patterns

Rhythm Patterns (RP) are a feature set for handling audio data based on analysis of the spectral audio data and psycho-acoustic transformations [14]. In a pre-processing stage, multiple channels are averaged to one, and the audio is split into segments of six seconds, possibly leaving out lead-in and fade-out segments, and further skipping other segments, e.g. out of the remaining segments every third may be processed.

The feature extraction process for a Rhythm Pattern is composed of two stages. For each segment, the spectrogram of the audio is computed using a Fast Fourier Transform. The window size is set to 1024 samples, applying a Hanning window of 50% overlap. The Bark scale groups frequencies to critical bands according to perceptive pitch regions, is applied to the spectrogram, aggregating it to 24 frequency bands. Then, the spectrogram is transformed into the decibel scale, and further psycho-acoustic transformations are applied. Subsequently, the values are transformed into the unit Sone, on which a doubling on the scale sounds to the human ear like a doubling of the loudness. This results in a psycho-acoustically modified representation reflecting human loudness sensation.

In the second stage, a discrete Fourier transform is applied to the Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band. After additional weighting and smoothing steps, a Rhythm Pattern exhibits magnitudes of modulation for 60 modulation frequencies on 24 bands, and has thus 1440 dimensions. Finally, the feature vectors of a songs segments are simply averaged by computing the median.

B. Statistical Spectrum Descriptors

Computing Statistical Spectrum Descriptors (SSD) features [5] relies on the first part of the algorithm for computing RP features, specifically on the Bark-scale representation of the frequency spectrum. From this representation of perceived loudness, seven statistical measures are computed for each of the 24 critical band, to describe fluctuations within them. The statistical measures comprise mean, median, variance, skewness, kurtosis, min- and max-value. A Statistical Spectrum Descriptor is extracted for each segment, and the SSD feature vector of a song is then calculated as the median of its segments. In contrast to the Rhythm Patterns feature set, the dimensionality of the feature space is much lower – SSDs have 168 instead of 1440 dimensions, still at matching performance in terms of genre classification accuracies [5].

C. Rhythm Histograms

The Rhythm Histogram [5] features are a descriptor for the rhythmic characteristics in a song. Contrary to the RP and the SSD, information is not stored per critical band. Rather, the magnitudes of each modulation frequency bin (at the end of the second stage of the RP calculation process) of all 24 critical bands are summed up, to form a histogram of 'rhythmic energy' for each of the 60 modulation frequencies. The RH feature vector for a piece of music is calculated as the median of the histograms of each segment. The dimensionality of RH of 60 features is much lower than with the other sets.

D. Bag-Of-Words

Classical bag-of-words indexing at first tokenises all text documents in a collection, most commonly resulting in a set of words representing each document. Let the number of documents in a collection be denoted by N, each single document by d, and a term or token by t. Accordingly, the *term frequency* $tf(t,d)$ is the number of occurrences of term t in document d and the *document frequency* $df(t)$ the number of documents term t appears in. The process of assigning weights to terms according to their importance or significance for a document is called 'term-weighing'. The weighing we rely on is the most common model, namely the *term frequency times inverse document frequency* [15], computed as $tf \times idf(t,d) = tf(t,d) \cdot ln(N/df(t))$. This results in vectors of weight values for each (lyrics) document d. In this representation it now is possible to compute similarities, as lyrics that contain a similar vocabulary are likely to be semantically related. We did not perform stemming in this setup, earlier experiments showed only negligible differences for stemmed and non-stemmed features [12]; the rationale behind using non-stemmed terms is the occurrence of slang language in some genres.

E. Rhyme Features

Rhyme denotes the consonance or similar sound of two or more syllables or whole words. The reason for considering rhyme as feature is that different genres of music should exhibit different styles of lyrics. We assume the rhyming characteristics of a song to be given by the degree and form of the rhymes used. 'Hip-Hop' or 'Rap' music, for instance, makes heavy use of rhymes, which (along with a dominant bass) leads to their characteristic sound. To identify such patterns we extract several descriptors from the song lyrics.

Our approach is based on a phoneme transcription of the lyrics. The words 'sky' and 'lie', for instance, both end with the same phoneme /ai/. The transcription is language dependent; however, our test collection is predominantly composed

TABLE I
RHYME FEATURES FOR LYRICS ANALYSIS

Feature Name	Description
Rhymes-AA	A sequence of two (or more) rhyming lines ('Couplet')
Rhymes-AABB	A block of two rhyming sequences of two lines ('Clerihew')
Rhymes-ABAB	A block of alternating rhymes
Rhymes-ABBA	A sequence of rhymes with a nested sequence ('Enclosing rhyme')
RhymePercent	The percentage of blocks that rhyme
UniqueRhymeWords	The fraction of unique terms used to build the rhymes

TABLE II
OVERVIEW OF TEXT STATISTIC FEATURES

Feature Name	Description
exclamation_mark, colon, single_quote, comma, question_mark, dot, hyphen, semicolon	simple counts of occurrences
d0 - d9	occurrences of digits
WordsPerLine	words / number of lines
UniqueWordsPerLine	unique words / number of lines
UniqueWordsRatio	unique words / words
CharsPerWord	number of chars / number of words
WordsPerMinute	the number of words / length of the song

of English tracks. After transcribing the lyrics to a phoneme representation, we distinguish two elements of subsequent lines in a song text: *AA* and *AB*. The former represents two rhyming lines, while the latter denotes non-rhyming. Based on these, we extract the rhyme patterns described in Table I. Subsequently, we compute the percentage of rhyming blocks, and define the unique rhyme words as the fraction of unique terms used to build rhymes, describing whether rhymes are frequently formed using the same word pairs. Experimental results indicate that more elaborate patterns based on assonance, semirhymes, or alliterations may well be worth studying.

F. Part-of-Speech Features

Part-of-speech (POS) tagging is a lexical categorisation or grammatical tagging of words. Different POS categories are e.g. nouns, verbs, articles or adjectives. We presume that different genres will differ also in the category of words they are using; thus, we extract several POS descriptors from the lyrics. We count the numbers of: *nouns*, *verbs*, *pronouns*, *relational pronouns* (such as 'that' or 'which'), *prepositions*, *adverbs*, *articles*, *modals*, and *adjectives*. To account for different document lengths, all of these values are normalised by the number of words of the respective lyrics document.

G. Text Statistic Features

Text documents can also be described by simple statistical measures based on word or character frequencies. Measures such as the average length of words or the ratio of unique words in the vocabulary might give an indication of the complexity of the texts, and are expected to vary over different genres. Further, the usage of punctuation marks such as exclamation or question marks may be specific for some genres, and some genres might make increased use of apostrophes when omitting the correct spelling of word endings. The list of extracted features is given in Table II.

All features that simply count character occurrences are normalised by the number of words of the song text to accommodate for different lyrics lengths. 'WordsPerLine' and 'UniqueWordsPerLine' describe the words per line and the unique number of words per line. The 'UniqueWordsRatio' is the ratio of the number of unique words and the total number of words. 'CharsPerWord' denotes the simple average number of characters per word. 'WordsPerMinute' is computed analogously to the well-known beats-per-minute (BPM) value.

IV. TEST COLLECTION

The collection we used in the following set of experiments was introduced in [16]. It is a subset of the collection marketed through the content download platform of Verisign Austria (http://www.verisign.at/), and comprises 60,223 of the most popular audio tracks by more than 7,500 artists. The collection contained a number of duplicate songs, which were removed for our experiments. For 41,679 songs, lyrics have been automatically downloaded from portals on the Web. We considered only songs that have lyrics with a certain minimum length, to remove lyrics that are most probably not correctly downloaded.

The tracks are manually assigned by experts to one or more of 34 different genres. 9,977 songs did not receive any ratings at all, and were thus not usable for our genre classification task. Further, we only kept songs that have a rather clear genre categorisation. Thus of those that received more than one voting, we only kept those that have at least two thirds of the votings agreeing on the same genre, removing 12,572 songs. Also, genres with less than 60 songs were not considered.

Finally, after all the removal steps, and thus considering only tracks that have both a clear genre assignment and lyrics in proper quality available, we obtain a collection of 20.109 songs, categorised into 14 genres. Details on the number of songs per genre can be found in Table III. It is noticeable that the different genres vary a lot in size. As such, the smallest class is "Classical", with just 62 songs, or 0.29%. Also, Scores / Soundtrack, Jazz, Blues, Dance / Electronic, Reggae and Comedy comprise less or just about 1% of the whole collection. Contrarily to this, the largest class, Pop, holds 6,156 songs, or 30.6%. Next are two almost equally sized classes, Alternative and Rock, each accounting for almost 3,700 songs or 18.4% of the collection. While this collection clearly is imbalanced towards Pop, Alternative Rock and Rock, accounting for more than two thirds of the collection, it can surely be regarded as a real-world collection. For the experimental results, the class distribution implies a baseline result of the size of the biggest class, thus 30.6%.

V. EXPERIMENTS

In our experiments we compare the performance of audio features and text features using various types of classifiers. We first extracted the audio and lyrics feature sets described

TABLE III
COMPOSITION OF THE TEST COLLECTION

Genre	Artists	Albums	Songs
Pop	1.150	1.730	6.156
Alternative	457	780	3.699
Rock	386	871	3.666
Hip-Hop	308	537	2.234
Country	223	453	1.990
R&B	233	349	1.107
Christian	101	170	490
Comedy	20	44	206
Reggae	30	48	121
Dance / Electronic	48	59	112
Blues	23	39	99
Jazz	38	49	97
Scores / Soundtrack	46	24	70
Classical	21	21	62
Total	3.084	5.074	20.109

in Section III, and then built several dozens of combinations of these different feature sets, both separately within the lyrics modality, as well as combinations of audio and lyrics feature sets. Most combinations are done with the SSD audio features, as this is the best performing audio set. For all experiments, we employed the WEKA machine learning toolkit [1], and unless otherwise noted used the default settings for the classifiers and tests. We used k-Nearest-Neighbour, Naïve Bayes, J48 Decion Trees, Random Forests, and Support Vector Machines. We performed the experiments based on a ten-fold cross-validation. All results given are micro-averaged classification accuracies. Statistical significance testing is performed per column, using a paired t-test with an α value of 0.05.

A. Single Feature Sets

Table IV gives an overview on the classification results. Generally, it can be observed that the results achieved with Naïve Bayes are extremely poor, and are below the above mentioned baseline of the percentage of the largest class, 30.61%, for all but one feature set. SVMs in contrast are performing best on most feature sets, except for those containing text statistics or POS features, where Random Forests are achieving the highest accuracies. In most cases, k-NN achieve better results when the parameter k is increased.

Regarding audio features, shown in the first section of Table IV, the highest accuracies are achieved with SSD features on all classifiers tested; all other feature sets are significantly inferior. Decision Trees on RP and RH features marginally fail to beat the baseline. k-NN seems to improve exceptionally well with a higher value of k – at k of 15, it almost equals the accuracies of the SVMs. Decision Trees and Random Forests perform poorly on RP, just matching the accuracies of RH features. Subsequently, we consider SSD, the highest performing feature set, as the objective we want to improve on in the following experiments on feature combinations.

For lyrics-based rhyme and style features shown in the second part of Table IV, the overall performance is not satisfying. Generally, the text-statistics features are performing best, with the exception of Naïve Bayes, which seems to have some problem with the data, ranking at 2.13%. Rhyme and POS features on SVMs achieve exactly the baseline, where the confusion matrix reveals that simply all instances have been classified into the biggest genre. The part-of-speech features with Naïve Bayes, and the text-statistics on SVM and the Decision Tree manage to marginally outperform the baseline, Decision Trees with statistical significance. Random Forests, however, perform quite well on the text statistics features, achieving almost 40%.

The third section in Table IV gives the results with the bag-of-words features, with different numbers of features selected via frequency thresholding (described in Section III-D). Compared to some of the audio-only features, the results are promising especially with SVMs, yielding the highest accuracy of 51.5%. The SVM classifier is known to perform well with high dimensionalities, which is clearly visible on the BOW features, where the accuracies with SVMs are increasing with the number of features are used. The results with SVMs clearly out-perform the RH features, and with a high dimensionality of 1,500 terms or more achieving even better results than RP; results with 2,000 and more terms show statistically significant improvements. With the other classifiers, the optimal number of features is varying, and not necessarily improving with more dimensions. The best result for k-NN is achieved with k=15 and 1,500 features, yielding 37.44%, clearly below results achieved with audio-only feature sets. With Decision Trees and Random Forests, the results are clearly better than with k-NN, and better than RH or RP features, with Decision Trees even better than the SSD.

Combining the rhyme and style features, slight improvements can be gained in most cases, as seen in the last section of Table IV. Besides Naïve Bayes, the best combination always includes the text statistics and part-of-speech features, and in two out of four cases also the rhyme features. However, the results are still far away from the lyrics features, and not that much better than the baseline. As for POS and text statistics features alone, the best results can be achieved with Random Forests, with around 40% in three of the four combinations.

B. Multi-Modal Feature Set Combinations

Even though the classification results of the lyrics-based features fall short of the SSD features, the objective to improve on, they can be combined to achieve statistically significant improvement over the audio-only results. With the big size of the dataset requiring many computational resources, we focused on combining the lyrics feature sets with SSDs, as they have clearly outperformed the other audio feature sets in the first set of experiments and our previous work.

The second section of Table V shows the results on combining SSD features with the rhyme and style features. It can be noted that each combination performs better than the SSD features alone, except on a few combinations on the k-NN classifier. For all combinations on SVM, most combinations on Decision Trees and Random Forests, as well as some combinations with the k-NN, the improvements are statistically

[1] http://www.cs.waikato.ac.nz/ml/weka/

TABLE IV
CLASSIFICATION ACCURACIES FOR SINGLE AUDIO AND TEXT FEATURE SETS, AS WELL AS RHYME AND STYLE FEATURE COMBINATIONS

Dataset	Dim.	1-NN	3-NN	5-NN	10-NN	15-NN	NB	DT	RF	SVM
RP	1440	38.69	39.41	43.56	45.60	46.13	17.25	30.35	37.72	49.47
RH	60	34.93	35.12	37.94	39.96	40.58	20.50	30.32	37.81	40.75
SSD	168	**46.26**	**46.87**	**49.55**	**51.61**	**52.16**	**26.45**	**39.42**	**48.69**	**55.77**
Rhyme	6	25.29	24.92	26.83	28.00	28.58	22.75	28.13	27.80	30.61
POS	9	27.96	26.61	30.10	31.88	32.78	**32.33**	27.81	32.74	30.61
TextStat	23	**29.40**	**28.32**	**31.56**	**33.27**	**33.57**	2.13	**33.62**	**39.14**	**32.39**
BOW_{20}	20	27.53	25.97	28.11	29.65	30.34	17.83	28.60	n/a	33.11
BOW_{50}	50	30.15	28.90	31.39	33.12	33.52	16.03	31.39	37.50	36.47
BOW_{200}	200	31.36	29.95	32.19	32.88	32.80	19.17	34.12	40.59	43.18
BOW_{399}	399	32.56	30.13	31.29	31.76	31.93	23.58	36.16	41.67	46.30
BOW_{798}	798	**33.20**	29.93	30.79	31.21	31.44	**23.79**	38.75	**42.21**	48.85
BOW_{896}	896	33.06	30.00	30.89	31.52	31.76	23.15	39.02	41.90	49.23
BOW_{997}	997	27.06	22.95	23.68	24.16	28.23	22.30	38.72	41.99	49.42
BOW_{1500}	1500	32.37	31.71	34.39	**36.44**	**37.44**	17.78	39.24	42.04	50.16
BOW_{2000}	2000	32.61	31.83	34.48	35.95	36.90	15.08	40.07	41.55	50.87
BOW_{2232}	2232	32.68	31.68	**34.51**	35.92	36.94	14.29	39.92	41.63	50.92
BOW_{2988}	2988	32.69	31.94	34.12	35.27	35.86	12.98	**41.13**	41.33	51.01
BOW_{3963}	3963	32.90	**32.08**	33.64	34.07	34.17	12.16	41.10	n/a	**51.50**
POS+Rhyme	15	27.91	26.87	29.31	31.08	32.35	**29.00**	27.97	34.33	30.59
POS+TextStat	32	**30.67**	29.66	**32.51**	**34.09**	**35.07**	3.48	**34.19**	**40.42**	35.22
Rhyme+TextStat	29	27.82	26.46	29.30	31.26	32.37	2.33	33.94	39.87	32.54
POS+Rhyme+TextStat	38	30.13	**29.68**	32.31	33.56	34.21	3.82	33.73	40.05	**36.09**

significant. The best result is achieved with SVMs when combining SSD with all of rhyme, part-of-speech and text statistics features. This combination achieves 58.09%, an improvement of 2.3 percent points over the baseline, with only a minor increase in the dimensionality. To offer a comparison on the combination performance, the third section in Table V shows results of combining SSD with RP, RH, and both of them. Only one combination on one classifier leads to significant improvements, while on some classifiers, some combinations even yield significant degradation compared to SSD features only. Also, the highest accuracy for each classifier is topped by several combinations with the rhyme and style features.

Combining SSD with the bag-of-words features, as seen in the third part of Table V, also leads to statistically significant improvements on Decision Trees, Random Forests, and SVMs, for the latter already with only 10 terms used. The best result is achieved on SVM when using around 1500 keyword dimensions with 60.71% classification accuracy, which is statistically significantly better than the SSD combination with the rhyme and style features. It can be noted that generally, k-NN do not profit from adding BOW features. On many combinations, especially with a smaller k, significant degradations can be observed. Only for a few combinations, with around 1,000 to 1,300 terms and higher k values, slight accuracy increased can be gained.

The last two parts of Table V finally present the results of combining all SSD, rhyme and style and bag-of-words features. One of these combinations also achieves the best result in this experiment series, namely SVMs on the last combination presented in the table, with 61.12%.

VI. CONCLUSION

We presented an evaluation on multi-modal features for automatic musical genre classification. Besides features based on the audio signal, we used a set of features derived from song lyrics as an additional, partly orthogonal dimension. Next to measuring the performance of single features sets, we in detail studied the power of combining audio and lyrics features.

Our main contribution in this paper is the large-scale of the evaluation of these features and their combination, on a database of over 20.000 songs. We showed that similar effects as for the smaller, carefully assembled databases of 600 and 3,100 songs presented in earlier work hold true as well for a larger database. Further, the database is taken from a real-world scenario, exhibiting potentially more challenging characteristics, such as having an imbalanced genre distribution.

One interesting observation is that the bag-of-words features alone already achieve very good results, even outperforming RP features, and not being far off the SSD. This, and the improved classification performance achieved on the combination of lyrics and audio feature sets, are promising results for future work in this area. Increased performance gains might be achieved by combining the different feature sets in a more sophisticated approach, e.g. by applying weighting schemes or ensemble classifiers.

REFERENCES

[1] J. Downie, *Annual Review of Information Science and Technology*. Medford, NJ: Information Today, 2003, vol. 37, ch. Music Information Retrieval, pp. 295–340.

[2] N. Orio, "Music retrieval: A tutorial and review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 1, pp. 1–90, September 2006.

[3] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[4] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organized Sound*, vol. 4, no. 30, pp. 169–175, 2000.

[5] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, London, UK, September 11-15 2005, pp. 34–41.

TABLE V

CLASSIFICATION ACCURACIES AND RESULTS OF SIGNIFICANCE TESTING FOR COMBINED AUDIO AND TEXT FEATURE SETS. STATISTICALLY SIGNIFICANT IMPROVEMENT OR DEGRADATION OVER DATASETS (COLUMN-WISE) IS INDICATED BY (+) OR (−), RESPECTIVELY

Dataset	Dim.	1-NN	3-NN	5-NN	10-NN	15-NN	DT	RF	SVM
SSD (test base)	168	46.26	46.87	49.55	51.61	52.16	39.42	48.69	55.77
SSD + POS	177	46.04	47.06	50.06	52.39	52.48	39.78	49.96 +	56.52 +
SSD + TextStat	191	46.94	47.70	50.58 +	52.22 +	52.95	42.46 +	51.13 +	57.35 +
SSD + Rhyme + TextStat	197	46.56	47.06	50.57	52.21	52.78	41.74 +	50.94 +	57.42 +
SSD + POS + TextStat	200	46.97	47.84 +	50.71 +	52.65	53.24 +	42.40 +	50.68 +	57.92 +
SSD + POS + Rhyme + TextStat	206	46.59	47.79 +	50.72 +	52.78 +	53.21 +	41.60	50.82 +	58.09 +
SSD + RP	1608	44.39 -	45.37 -	48.60	50.75	51.53	35.69 -	43.63 -	55.77
SSD + RH	228	46.75	47.19	50.24	52.19	52.77	39.12	48.96	57.42 +
SSD + RP + RH	1668	44.38 -	45.62 -	48.69	50.59	51.45	36.51 -	43.59 -	55.60
BOW_{10} + SSD	178	39.87 -	40.20 -	43.90 -	46.13 -	46.30 -	39.45	48.78	56.32 +
BOW_{20} + SSD	188	40.11 -	40.83 -	44.95 -	46.95 -	48.00 -	39.38	49.14	57.12 +
BOW_{50} + SSD	218	41.52 -	43.42 -	46.68 -	48.71 -	49.70 -	39.54	49.36	58.46 +
BOW_{75} + SSD	243	41.86 -	43.45 -	46.21 -	49.01 -	49.56 -	39.78	48.74	58.84 +
BOW_{150} + SSD	318	41.98 -	43.44 -	46.64 -	48.55 -	49.80 -	39.10	49.16	59.17 +
BOW_{300} + SSD	468	42.05 -	42.77 -	45.22 -	48.10 -	50.00 -	40.82 +	48.62	59.98 +
BOW_{798} + SSD	966	39.50 -	42.46 -	47.33 -	50.52	51.64	41.72 +	48.32	60.49 +
BOW_{997} + SSD	1165	42.33 -	44.69 -	49.08	51.52	52.47	42.01 +	48.05	60.69 +
BOW_{1248} + SSD	1416	43.91 -	45.92	49.58	51.57	52.15	42.26 +	48.37	60.51 +
BOW_{1500} + SSD	1668	44.15 -	45.61 -	49.29	50.80	51.59	42.22 +	47.70 -	60.71 +
BOW_{2232} + SSD	2400	43.13 -	45.10 -	48.18 -	49.68 -	49.55 -	42.84 +	47.01 -	60.70 +
BOW_{2988} + SSD	3156	42.02 -	43.87 -	46.96 -	47.91 -	47.99 -	43.52 +	47.33 -	60.41 +
BOW_{3963} + SSD	4131	41.45 -	43.01 -	45.17 -	46.79 -	46.73 -	43.36 +	n/a	60.67 +
BOW_{10} + SSD + TextStat	201	40.46 -	41.04 -	44.28 -	46.29 -	46.73 -	42.09 +	50.70 +	57.59 +
BOW_{20} + SSD + TextStat	211	40.61 -	41.54 -	44.91 -	47.63 -	48.52 -	42.75 +	50.69 +	58.14 +
BOW_{50} + SSD + TextStat	241	42.40 -	43.89 -	47.12 -	49.04 -	49.90 -	42.54 +	50.07	59.16 +
BOW_{150} + SSD + TextStat	341	43.00 -	44.19 -	47.22 -	49.43 -	50.47 -	41.26 +	50.35 +	59.95 +
BOW_{399} + SSD + TextStat	590	41.99 -	42.55 -	45.11 -	49.16 -	51.00	41.01 +	50.25 +	60.55 +
BOW_{799} + SSD + TextStat	989	40.80 -	44.11 -	48.64	51.21	52.13	41.51	49.34	60.68 +
BOW_{997} + SSD + TextStat	1188	45.26	46.76	49.93	51.86	52.90	41.63 +	49.10	60.54 +
BOW_{1248} + SSD + TextStat	1439	44.30 -	46.43	49.83	52.26	52.80	42.10 +	48.91	60.66 +
BOW_{1500} + SSD + TextStat	1691	44.23 -	46.22	49.54	51.79	52.68	42.05 +	48.37	60.87 +
BOW_{2232} + SSD + TextStat	2423	43.73 -	45.83	48.80	50.34	50.73 -	42.93 +	47.86	60.76 +
BOW_{2988} + SSD + TextStat	3179	42.38 -	44.49 -	47.35 -	48.47 -	48.73 -	43.32 +	47.40	60.63 +
BOW_{3963} + SSD + TextStat	4154	41.81 -	43.06 -	46.00 -	47.40 -	47.23 -	43.54 +	n/a	61.05 +
BOW_{10} + SSD + TextStat + POS + Rhyme	216	41.05 -	41.91 -	45.11 -	47.00 -	47.40 -	41.60	50.65 +	58.39 +
BOW_{20} + SSD + TextStat + POS + Rhyme	226	41.71 -	42.41 -	45.95 -	47.80 -	48.79 -	42.36 +	50.14 +	58.90 +
BOW_{50} + SSD + TextStat + POS + Rhyme	256	43.26 -	44.63 -	47.77 -	49.83 -	50.66 -	42.24 +	50.51 +	59.62 +
BOW_{150} + SSD + TextStat + POS + Rhyme	356	44.64 -	45.49 -	48.61	50.69 -	51.52 -	40.78	50.29 +	60.07 +
BOW_{399} + SSD + TextStat + POS + Rhyme	605	43.96 -	44.45 -	47.90 -	51.09	52.96	41.09	50.23 +	60.72 +
BOW_{798} + SSD + TextStat + POS + Rhyme	1004	43.22 -	45.80	50.28	52.57	53.56 +	41.58 +	48.83	60.67 +
BOW_{997} + SSD + TextStat + POS + Rhyme	1203	45.68	47.70 +	50.98 +	53.39 +	53.78 +	41.45 +	48.91	60.86 +
BOW_{1248} + SSD + TextStat + POS + Rhyme	1454	44.88 -	47.40	50.96 +	53.05 +	53.76 +	42.38 +	48.71	61.06 +
BOW_{2232} + SSD + TextStat + POS + Rhyme	2438	43.82 -	46.44	49.38	51.68	52.13	43.02 +	47.29 -	60.92 +
BOW_{2988} + SSD + TextStat + POS + Rhyme	3194	42.85 -	44.85 -	48.27 -	49.59 -	49.98 -	43.23 +	47.58	60.79 +
BOW_{3963} + SSD + TextStat + POS + Rhyme	4169	41.87 -	43.72 -	46.66 -	48.13 -	48.34 -	43.48 +	n/a	61.12

[6] J. P. G. Mahedero, Á. Martínez, P. Cano, M. Koppenberger, and F. Gouyon, "Natural language processing of lyrics," in *Proceedings of the ACM 13th International Conference on Multimedia (MM'05)*, New York, NY, USA, 2005, pp. 475–478.

[7] B. Logan, A. Kositsky, and P. Moreno, "Semantic analysis of song lyrics," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'04)*, Taipei, Taiwan, June 27-30 2004, pp. 827–830.

[8] D. Yang and W. Lee, "Disambiguating music emotion using software agents," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

[9] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," San Diego, CA, USA, December 11–13 2008, pp. 688–693.

[10] S. Baumann, T. Pohle, and S. Vembu, "Towards a socio-cultural compatibility of mir systems." in *Proceedings of the 5th International Conference of Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October 10-14 2004, pp. 460–465.

[11] E. Brochu, N. de Freitas, and K. Bao, "The sound of an album cover: Probabilistic multimedia and IR," in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, C. M. Bishop and B. J. Frey, Eds., Key West, FL, USA, January 3-6 2003.

[12] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, September 14-18 2008, pp. 337–342.

[13] ——, "Combination of audio and lyrics features for genre classification in digital audio collections," in *Proceedings of the ACM Multimedia 2008*, Vancouver, BC, Canada, October 27-31 2008, pp. 159–168.

[14] A. Rauber, E. Pampalk, and D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles," in *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'02)*, Paris, France, October 13-17 2002, pp. 71–80.

[15] G. Salton, *Automatic text processing – The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[16] F. Kleedorfer, P. Knees, and T. Pohle, "Oh oh oh whoah! towards automatic topic detection in song lyrics," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA, USA, September 14 – 18 2008, pp. 287 – 292.

metamidi: a tool for automatic metadata extraction from MIDI files

Tomás Pérez-García, José M. Iñesta, and David Rizo
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante
Alicante, 03080, Spain
e-mail: {tperez,inesta,drizo}@dlsi.ua.es

***Abstract*—The increasing availability of on-line music has motivated a growing interest for organizing, commercializing, and delivering this kind of multimedia content. For it, the use of metadata is of utmost importance. Metadata permit organization, indexing, and retrieval of music contents. They are, therefore, a subject of research both from the design and automatic extraction approaches. The present work focuses on this second issue, providing an open source tool for metadata extraction from standard MIDI files. The tool is presented, the utilized metadata are explained, and some applications and experiments are described as examples of its capabilities.**

I. Introduction

Metadata permit organization, indexing, and retrieval of music contents in digital collections, e.g. digital libraries [10] or on-line music stores, to name but a few. A digital music library requires addressing complex issues related to description, representation, organization, and use of music information [9]. Another application of accurate metadata is related to copyright issues when sharing of music databases among music researches located in different countries. A suitable solution is the use of metadata instead of the source song files. The presented tool along with its helper formats can help in that direction.

Metadata are, therefore, an important subject of research focusing on both the design and automatic extraction approaches. The quality of content-based music recommendations is importantly influenced by the number and quality of available metadata attributes [3].

A number of efforts have been addressed to automatic metadata extraction from music data, both from audio and symbolic files. For example, MPEG-7 [6] deals with multimedia content description. In this context, Pachet in [11] described a project on the automatic computation of those descriptors focusing on digital audio data, while Kleedorfer et al. [4] aimed at generation of metadata attributes automatically extracted from song lyrics. Some of these metadata are utilized, through standard text classification methods, to derive genre, mood, and offensive content tags.

The present work focuses on metadata extraction from standard MIDI files, introducing `metamidi`, an open source tool for metadata extraction from them. A number of works in the literature have deal with related problems, reporting MIR platforms for symbolic file format analysis in different languages like Matlab [2] or java [8]. These works are able to analyze technical properties of music formats, ranging from Humdrum, Lilypond ([5]) or MIDI ([2], [8]). Some of these works are based on solving a music information retrieval (MIR) problem, like genre extraction from content, while others are application-independent. Our proposal falls in the latter category, focusing on metadata like text metaevents, key, tempo, and meter signatures, etc., although some statistical properties based on track contents are also provided. This range of data, permits to address organization oriented problems, like categorization and indexation of MIDI files, but also well know MIR problems like melody part or genre recognition.

A special attention has been paid to interfacing with other programs for metadata analysis, so the software has been designated to operate in command line.

The rest of the paper is structured as follows. The obtained metadata are explained in section II, the `metamidi` tool is presented in section III, and some applications and experiments are described in section IV as examples of its capabilities.

II. MIDI metadata

`metamidi` extracts descriptive, structural, and technical metadata at two different levels: global and track-local metadata. Output is provided in three different formats:

- register format, where metadata are presented in a report style useful for readability,
- line format, oriented to communication with other software for further processing, and
- XML format, providing a standard and modern format for interfacing with other applications. The use of XML-based format has proven to be useful in former works [1].

More details on these formats are provided below in section III.

A. Description of global metadata

In this section the general metadata extracted from a MIDI file, are described. These metadata are extracted from the file storage system, from its header, and from its sequence tracks.

- **Name**, the complete path where the MIDI file is stored.
- **Metaeventtext**, strings contained in text metaevents. This information includes all kind of text metaevents found

in the track 0 of a MIDI file, like author, song name, copyright notices, sequencer, etc.
- **Size**, the size in bytes of the file.
- **Format**, the format of a MIDI file. 0 for single track; 1 for multi-track files, and 2 for multi-sequence files.
- **Number of tracks**, the number of tracks that the file is composed of.
- **Resolution**, the number of ticks per beat.
- **Tempo**, initial value for tempo in beats per minute.
- **Tempo changes**, the number of tempo change events. Tempo tracking has not been implemented, due to the large number of changes that may appear if *accelerandos* or *rittardandos* are present in the sequence.
- **Meter**, the list of number of beats per bar over beat kind separated by commas. The pulse where it changes is showed between brackets.
- **Meter changes**, the number of meter changes.
- **Key**, the list of tonality metaevent in the file. The pulse where it changes is showed between brackets.
- **Key changes**, the number of tonality changes.
- **Instruments**, the numbers of the General MIDI patches utilized in the sequence.
- **Percussion**, the percussion instruments utilized in the sequence according to the pitch values utilized in channel 10, assuming the General MIDI standard percussion map. Percussion instruments have been categorized in three different groups, coded as follows:
  - 1 : instruments usually found in a drum kit,
  - 2 : latin percussion, and
  - 3 : other percussion elements.
- **Sequence duration**, the number of clock ticks where the offset of the last note happens.
- **Has sysex**, a flag showing whether sysex messages appear (value 1) in the sequence or not (value 0).

B. Description of each track metadata:

The second level of metadata are descriptions of each track content. Therefore, the metadata described below are for each track.

- **Metaeventtext**, strings contained in text metaevents included track, like track name, instrument name, lyrics, markers, etc.
- **Track duration**, the number of ticks from the onset of the first note to the offset of the last note in the track.
- **Duration rate**, = track duration / sequence duration.
- **Occupation**, the sum of the number of ticks where notes are sounding in the track.
- **Occupation rate**, = occupation / track duration.
- **Polyphony duration rate**, ticks where two or more notes are sounding / occupation.
- **max polyphony**, the number of maximum simultaneous notes in the track.
- **Avg polyphony**, number of sounding notes in average (weighted by their durations), computed as

$$= \frac{\sum_{\forall n>0} n \times (\#\text{ticks with } n \text{ notes})}{\text{occupation}} \qquad (1)$$

- **Low pitch**, lowest pitch in the track.
- **High pitch**, highest pitch in the track.
- **Modulations**, number of modulation messages.
- **Aftertouches**, number of aftertouch messages.
- **Pitch bends**, number of pitch alteration messages.
- **Program changes**, patch change messages. The pulse where it changes is showed between brackets.

III. The `metamidi` tool

`metamidi` can be freely downloaded from `http://grfia.dlsi.ua.es/gen.php?id=resources`. In the downloading page the reader can find the source code of `metamidi`, the DTDs listed in Table I, and an alternative XSD format [1] that can help in the automatic developing of automatic parsers like the one programmed in Java with XMLBeans [2] that is also provided.

It has been developed in ANSI C 4.2.4 version and tested under Linux. It is designed to operate in the command line, providing the different output formats, with this syntax:

```
metamidi -{r|x|l} file [-o fileoutput]
```

where:

- `-r` : outputs metadata in register format.
- `-l` : outputs metadata in line format. A "|" separates the global metadata from track metadata and also the different track metadata parts. A ";" character is used to separate each feature. "," character separates multi-valuated features.
- `-x` : outputs metadata in XML format. A DTD is provided where tags and attributes are described (see Table I)

When a string metadata is not found, a "$" is given. For missing numerical values, a -1 is displayed.

If no output file is provided, the standard output is utilized, providing a way to pipe the result to other data processing software that uses the standard input as input. This permits to design powerful scripts using off-the-shelf text and data processing utilities.

In table II an example of a two-track MIDI file is displayed when `metamidi` operates in report format.

IV. Application examples

The kind of extracted metadata can be used in a number of applications like MIDI file indexing, organizing, classifying, etc. In this section, two well known MIR problems have been addressed using the provided metadata. Firstly, we will report results on selecting the track containing the melody in a multi-track MIDI file, and secondly, we will present results of an experiment on genre classification using the timbral metadata provided by the patch instrument map in the file.

[1] http://www.w3.org/XML/Schema
[2] http://xmlbeans.apache.org/

TABLE I
PROPOSED DTD

```
<!ELEMENT midifile (external,global,tracks)>
<!ELEMENT external (comments?)>
<!ATTLIST external
  name CDATA #REQUIRED
  size CDATA #REQUIRED
  midiformat (0|1|2) #REQUIRED
  numtracks CDATA #REQUIRED
  resolution CDATA #REQUIRED
>
<!ELEMENT comments (#PCDATA)>
<!ELEMENT global (comments?)>
<!ATTLIST global
  metaeventtext CDATA #REQUIRED
  tempo CDATA #REQUIRED
  tempoChanges CDATA
  meter CDATA #REQUIRED
  meterChanges CDATA
  key CDATA #REQUIRED
  keyChanges CDATA
  instruments CDATA
  percussion (-1|1|2|3)
  duration CDATA #REQUIRED
  hasSysEx (true|false) #REQUIRED
>
<!ELEMENT tracks (track+)>
<!ELEMENT track (comments?)>
<!ATTLIST track
  metaeventtext CDATA #REQUIRED
  channel CDATA #REQUIRED
  duration CDATA #REQUIRED
  durationRate CDATA #REQUIRED
  occupation CDATA #REQUIRED
  occupationRate CDATA #REQUIRED
  polyphonyDurationRate CDATA #REQUIRED
  maxPoliphony CDATA #REQUIRED
  avgPoliphony CDATA #REQUIRED
  low pitch CDATA #REQUIRED
  high pitch CDATA #REQUIRED
  modulations CDATA
  afterTouches CDATA
  pitchBends CDATA
  programChanges CDATA
>
```

TABLE II
OUTPUT EXAMPLE OF A TWO TRACK MIDI FILE USING THE "REPORT" FORMAT.

```
name:              /home/repository/Beethoven/Fur-Elise.mid
text metaevent:          Fur Elise,Ludwig van Beethoven
size:                    9574
format:                  1
num tracks:              3
resolution:              480
tempo:                   75.00
tempo changes:           25
meter:                   4/4(0),3/8(0),3/8(5760)
meter changes:           3
key:                     CM(0)
key changes:             1
instruments:             1
percussion:              -1
duration:                90000
has sysex:               0
------ Features of track 1 ----
text metaevent:          Piano RH
channel:                 1
duration:                89968
duration rate:           1.00
occupation:              75348
occupation rate:         0.84
polyphony duration rate: 0.20
max polyphony:           4
avg polyphony:           1.30
low pitch:               57
high pitch:              100
modulations:             0
aftertouches:            0
pitch bends:             0
program changes:         1(0)
------ Features of track 2 ----
text metaevent:          Piano LH
channel:                 2
duration:                90000
duration rate:           1.00
occupation:              39587
occupation rate:         0.44
polyphony duration rate: 0.19
max polyphony:           3
avg polyphony:           1.23
low pitch:               33
high pitch:              76
modulations:             0
aftertouches:            0
pitch bends:             0
program changes:         1(960)
```

A. Melody track selection

Standard MIDI files are structured as a number of tracks. One of them usually contains the melodic line of the piece, while the other tracks contain accompaniment music. Finding that melody track is very useful for a number of applications, including music retrieval or motif extraction, among others. In [12] the authors introduced a method to identify the track that contains the melody using statistical properties of the musical content and pattern recognition techniques.

Here we are going to use metadata extracted through `metamidi` to perform the same task under a Gaussian approach. For that, we can compute the probability of a track to contain the melody of a MIDI file from different track metadata:

- the amount of music information in the track (occupation rate),
- melody are usually monophonic, so the use of polyphony is important (polyphony duration rate, max polyphony, average polyphony),
- for a melody to be sung it must be *cantabile*, so the pitch ranges are relevant (low pitch and high pitch).

In the training phase, given one of the former metadata, d, its probability distribution function is computed assuming a Gaussian distribution, both for the metadata values of the tracks labeled as melodies, obtaining $P(d|M)$, and for those extracted from non melody tracks, obtaining $P(d|\neg M)$.

During the test phase, in order to select the melody track from a midi file, the a posteriori probabilities for all the tracks are computed using the Bayes theorem:

$$P(M|d_i) = \frac{P(d_i|M)P(M)}{P(d_i|M)P(M) + P(d_i|\neg M)P(\neg M)} \qquad (2)$$

where i subindex refers to the values extracted from the i-th track. The a priori probability for a track of being a melody, $P(M)$ is computed from all the files in the training set as number melody tracks / total number of tracks, and $P(\neg M) =$

$1 - P(M)$.

The final decision is taken using a maximum likelihood decision taking into account those tracks with probabilities higher than a threshold, θ:

$$\hat{t}_M = \arg\max_i \{P(M|d_i) > \theta\} \quad (3)$$

If no track has a $P(M|d_i) > \theta$, the decision of "no melody" is given for that file.

There is also the possibility of combining two probability distributions. The same methodology is applied, but in this case, the two Gaussians involved are multiplied, so $P(d|M) = P(d_A|M) \times P(d_B|M)$, and therefore, the same has to be done with the respective decision thresholds, $\theta = \theta_A \times \theta_B$. The rest of equations remain the same.

The same 6 data sets previously utilized in [12] have been utilized for the experiments. Midi files from pop-rock ("Kar"), jazz ("Jazz"), and classical music ("Clas") are organized in the different data sets in order to test the specificities of this problem depending on the music genre. The sets with a "200" suffix are smaller sets, more uniform in their structure, while the other three are bigger ones, more heterogeneous and, therefore, more difficult for their melody track to be identified. All of them have been manually tagged and multiple melody track and no melody track situations occur.

The system is evaluated as follows. Defining TP as the number of true positive decisions, FP as the number of false positive decisions, TN the number of true negative decisions, and FN as the number of false negative decisions, the evaluation parameters were:

$$\text{Success}: \quad S = 100\frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

$$\text{Precision}: \quad P = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall}: \quad R = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F}-\text{measure}: \quad F = \frac{2RP}{R+P} \quad (7)$$

The figures presented in Table III are the results of a 10-fold cross-validation, where 10 sub-experiments were performed, using 9/10 of the set for training, keeping 1/10 for testing. The results presented are the best obtained for each data set, including the metadata they were obtained with and the utilized threshold.

TABLE III
RESULTS OF MELODY TRACK IDENTIFICATION FOR THE DIFFERENT DATA SETS.

Corpus	Descr.	θ	S%	P	R	F
Clas200	High p.	0.05	98.5	0.99	1.00	0.99
Clas	Avg.Poly	0.20	80.8	0.81	1.00	0.89
Jazz200	Max.Poly & Low p.	0.1^2	86.5	0.88	0.98	0.93
Jazz	Max.Poly & Low p.	0.05^2	82.0	0.83	0.99	0.90
Kar200	Avg.Poly	0.10	80.9	0.81	1.00	0.89
Kar	Avg.Poly	0.10	88.6	0.87	1.00	0.94

B. Timbre-based genre classification

The timbral information provided by `metamidi` can be utilized to infer the genre of the music the MIDI file contains. In fact, there are examples in the literature of using the set of instrument patch numbers (together with other descriptors) to classify MIDI files into music genres [7]. In that work, instrument patch numbers are coded in a vector and a distance-based method to instrument class vectors is used to classify. Nevertheless, the approach proposed here is a probabilistic one, in which a MIDI file has a probability of being of a given genre depending on the probabilities of its instruments to be used in that genre.

We have a set of classes $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ and a labeled training set of songs, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{|\mathcal{X}|}\}$. A MIDI file is represented as a vector $\mathbf{x}$, where each component $x_i \in \{0, 1\}$ codes the absence or presence of the patch t in the *instruments* and *percussion* metadata. The dimensionality of the vectors is $D = 131$, corresponding to the 128 General MIDI patches plus the 3 percussion categories described in section II. A given MIDI file is assigned to the class $c_j \in \mathcal{C}$ with maximum a posteriori probability:

$$P(c_j|\mathbf{x}) = \frac{P(c_j)P(\mathbf{x}|c_j)}{P(\mathbf{x})} \quad , \quad (8)$$

where $P(c_j)$ is the a priori probability of class c_j computed in this case as $P(c_j) = 1/|\mathcal{C}|$, assuming that all genres are equally probable, $P(\mathbf{x}) = \sum_{j=1}^{|\mathcal{C}|} P(c_j)P(\mathbf{x}|c_j)$ is a normalization factor, and $P(\mathbf{x}|c_j)$ is the probability of $\mathbf{x}$ being generated by class c_j given by a multivariate Bernoulli distribution of instruments in class c_j, learned from the training set:

$$P(\mathbf{x}|c_j) = \prod_{i=1}^{D} x_i P(t_i|c_j) + (1 - x_i)(1 - P(t_i|c_j)) \quad (9)$$

where $P(t_i|c_j)$ are the class-conditional probabilities of each patch, t_i, in the instrument map, that can be easily calculated by counting the number of occurrences of each instrument in the corresponding class:

$$P(t_i|c_j) = \frac{1 + M_{ij}}{2 + M_j} \quad (10)$$

where M_{ij} is the number of files in class c_j containing the instrument t_i, and M_j is the total number of songs in class c_j. This equation permits to avoid zero probabilities when a previously unseen instrument is found in the test file.

A data base of 856 midi files from popular, jazz, and academic music has been utilized. Popular music data have been separated into three sub-genres: *pop*, *blues*, and *celtic* (mainly Irish jigs and reels). For jazz, three styles have been established: a *pre-bop* class grouping swing, early, and Broadway tunes, *bop* standards, and *bossanovas* as a representative of latin jazz. Finally, academic music has been categorized according to historic periods: *baroque*, *classicism*,

and *romanticism*. This data base is available upon request to the authors.

The input metadata for this application have been easily obtained through the following command (syntax needs to be adapted to the used console language):

```
metamidi -l *.mid | cut -d ";" -f 13,14
```

positions 13 and 14 correspond to *instruments* and *percussion* metadata, respectively, in the output line.

Tables IV and V show the confusion matrices for the two experiments performed, respectively, grouping the files in the three broad music categories (popular, jazz, and academic) and in the nine classes described above. Rows are the ground-truth classes and columns are the system predictions. Both experiments were designed following a 10-fold cross validation scheme.

TABLE IV
CONFUSION MATRIX FOR THE THREE GENRE RECOGNITION TASK.

	Academic	Jazz	Popular
Academic	228	6	1
Jazz	3	295	40
Popular	5	16	262

The overall performance for the three-classes classification was 93 ± 2, showing the good performance of timbral metadata. For the nine-classes problem, a much harder one, the overall classification success was 68 ± 5. Note that a by-chance classifier is expected to achieve 39.5 and 20.8%, respectively, corresponding en each case to the most numerous class, so the improvements are 53.5 and 47.2%. Also it is important to see that most of the errors occur now among genres of the same broad category, like classical and romantic music, or pre-bop and bop songs.

TABLE V
CONFUSION MATRIX FOR THE NINE GENRE RECOGNITION TASK.

	bar	clas	rom	pre	bop	bos	cel	blu	pop
baroque	35	5	16	0	0	0	0	0	0
classic	8	1	41	0	0	0	0	0	0
romantic	12	7	110	0	0	0	0	0	0
prebop	0	0	0	150	22	6	0	0	0
bop	0	0	0	78	10	6	0	0	0
bossa	0	0	0	3	1	52	1	0	9
celtic	0	0	0	2	0	1	96	0	0
blues	0	0	0	7	1	1	5	62	8
pop	0	0	0	1	2	13	4	7	73

V. CONCLUSION

Metadata are achieving a growing interest in music information retrieval applications thanks to the high level information they provide. The development of systems for automatically extract them from digital files is one of the main concerns. In this paper, `metamidi` has been presented as an open source software for descriptive, structural, and technical metadata extraction from standard MIDI files. These metadata are extracted from both file properties and track properties.

An illustration of its performance has been presented addressing two MIR problems through the use of the provided metadata: melody track identification and music genre recognition. The good performances obtained show the power of metadata automatic extraction tools for organizing, indexing, and accessing music information in the context of multimedia digital libraries.

ACKNOWLEDGMENT

This work is supported by the Spanish Ministry projects: TIN2006–14932–C02 and Consolider Ingenio 2010 (MIPRCV, CSD2007-00018), both partially supported by EU ERDF.

REFERENCES

[1] A. Baratè, G. Haus, and L. A. Ludovico, *Music representation of score, sound, MIDI, structure and metadata all integrated in a single multilayer environment based on XML*. Hershey, PA: Idea Group Reference, 2007.

[2] T. Eerola and P. Toiviainen, "Mir in matlab: The midi toolbox," in *ISMIR*, 2004, pp. 22–27.

[3] F. Kleedorfer, U. Harr, and B. Krenn, "Making large music collections accessible using enhanced metadata and lightweight visualizations," in *Proceedings of the 3rd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS '07)*, Barcelona, Spain, November 2007, pp. 138–144.

[4] F. Kleedorfer, P. Knees, and T. Pohle, "Oh oh oh whoah! towards automatic topic detection in song lyrics," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, September 2008, pp. 287–292.

[5] I. Knopke, "The perlhumdrum and perllilypond toolkits for symbolic music information retrieval," in *Proceedings of the 2008 International Conference on Music Information Retrieval*, 2008, pp. 147–152.

[6] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG7: Multimedia Content Description Interface*. West Sussex, England: John Wiley & Sons, 2002.

[7] C. McKay and I. Fujinaga, "Automatic genre classification using large high-level musical feature sets," in *Proc. of the 5th International Conference on Music Information Retrieval, ISMIR 2004*, 2004, pp. 525–530.

[8] C. Mckay and I. Fujinaga, "jsymbolic: A feature extractor for midi files," in *In Int. Computer Music Conf*, 2006, pp. 302–305.

[9] N. Minibayeva and J. W. Dunn, "A digital library data model for music," in *Proc. of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, Oregon, 2002, pp. 154–155.

[10] M. Notess and J. Dunn, "Variations2: improving music findability in a digital library through work-centric metadata," in *JCDL'04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. Tucson, AZ: ACM Press, 2004, p. 422.

[11] F. Pachet, "Metadata for music and sounds: The Cuidado project," in *Proceedings of the CBMI Workshop*, University of Brescia, september 2001, pp. 411–415.

[12] D. Rizo, P. J. P. de León, C. Pérez-Sancho, A. Pertusa, and J. M. Iñesta, "A pattern recognition approach for melody track selection in midi files," in *Proc. of the 7th Int. Symp. on Music Information Retrieval ISMIR 2006*, T. A. Dannenberg R., Lemström K., Ed., Victoria, Canada, 2006, pp. 61–66.

Integration of Chroma and Rhythm Histogram Features in a Music Identification System

Riccardo Miotto and Nicola Montecchio
Department of Information Engineering
University of Padova
Padova, Italy
{riccardo.miotto,nicola.montecchio}@dei.unipd.it

***Abstract*—A system for Music Identification is presented which makes use of two different feature descriptors, namely Chroma and Rhythm Histogram features. A simple approach to feature combination is proposed and evaluated against a test collection. Finally directions for future work are proposed, focusing on performance optimization towards a scalable system.**

I. Introduction

The increasing availability of large music collections poses challenging research problems regarding the organization of documents according to some sense of similarity. Following a peculiar social phenomenon in the last years, an increasing number of users is joining social communities to upload their personal recordings and performances. Content-based music identification has become an important research topic because it can provide tools to efficiently retrieve and organize music documents. In particular, the large availability of non-commercial recordings puts a major interest towards the *cover identification* problem. Generally, the term *cover* defines a new rendition of a previously recorded song in genres such as rock and pop. Cover songs can be either live or studio recordings and may have a completely different arrangement.

An earlier approach to music identification was *audio fingerprinting*, that consists in a content-based signature of a music recording to describe digital music even in presence of noise, distortion, and compression [1]. On the contrary, cover identification approaches must be able to identify a song from the recording of a performance, yet independently from the particular performance. For example, identification of live performances may not benefit from the fingerprint of other performances, because most of the acoustic parameters may be different. Collecting all the possible live and cover versions of a music work is clearly unfeasible.

Cover music identification methodologies described in literature generally exploit the well-known *Chroma* features to describe the harmonic content of the music recordings. In particular they have been widely exploited in [2], [3] and [4].

Since Chroma features are high dimensional, they considerably affect computational time for search operations, especially considering the management of different tempo with alignment techniques. Efficiency becomes then a key issue if an identification system is proposed to a large community of users, as in the case of a Web-based music search engine.

In [4], we proposed an efficient methodology to identify classical music recordings by applying the Locality Sensitive Hashing (LSH) paradigm [5], a general approach to handle high dimensional spaces by using ad-hoc hashing functions to create collisions between vectors that are close in the space. LSH has been applied to efficient search of different media [6].

The focus of this paper is on the integration of feature descriptors that are relative to two characterizing aspects of a song: *harmonic* and *rhythmic* content. The main idea is that usually cover songs preserve not only the harmonic-melodic characteristics of the original work but also its rhythmic profile. Starting from the idea proposed in [4], we propose a cover identification system which combines the efficient hash-based Chroma descriptors with a rhythmic profile descriptor in order to increase the identification precision.

II. System model

The system works by combining the evidence given by Chroma and Rhythmic descriptors into a single ranking of possible matching songs. The rhythm descriptors used are *Rhythm Histogram* (RH) features [7], which were originally proposed in a genre classification task.

While Chroma features have been thoroughly investigated previously and are provided with an efficient implementation, RH features have only recently been adopted by us and their performances in terms of speed are not comparable yet; both aspects are described below. An overview of the system is depicted in Figure 1.

A. Chroma features

A music descriptor widely applied to cover identification is Chroma features. As is well-known, Chroma features are related to the intensity associated with each of the 12 semitones within an octave, with all octaves folded together. The concept behind chroma is that the perceived quality of a chord depends only partially on the octaves in which the individual notes are played. Instead, what seems to be relevant are the pitch classes of the notes (the names of the notes on the chromatic scale) that form a chord. This robustness to changes in octaves is also exploited by artists who play a cover song: while the main melody is usually very similar to the original one, the accompaniment can have large variations without affecting the recognizability of the song.

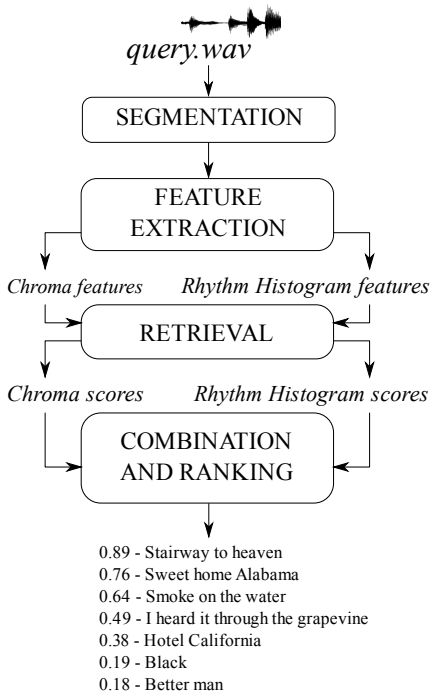


Fig. 1. Overview of the system model

As described in [4], a Chroma vector $\mathbf{c}$ is a 12-dimensional vector of pitch classes, computed by processing a windowed signal with a Fourier transform. According to the approach proposed in [8], chroma features have been computed using the instantaneous frequency within each FFT bin to identify strong tonal components and to achieve higher resolution. In Figure 2(a) a Chroma vector corresponding to an A7 chord is depicted; Figure 2(b) shows the evolution of Chroma vectors over time for an excerpt of the song "Heroes" by D. Bowie.

For each vector $\mathbf{c}$, quantization $\mathbf{q}$ is achieved by considering the ranks of the chroma pitch classes, instead of their absolute values, to obtain a general representation robust to variations due to different performing styles. In particular, rank-based quantization is carried out by computing the rank of the value of the energy in the various pitch classes.

Rank-based quantization aims at a final compact representation, which can be obtained by considering that a vector $\mathbf{q}$ can be thought as a twelve digit number represented in base k. A simple hashing function h can be computed by obtaining the decimal representation of this number, according to equation

$$h = \sum_{i=1}^{12} k^{i-1} \mathbf{q}_i \qquad (1)$$

where additional hashing techniques can be applied to store the values h in one array, which can be accessed in constant time. A typical technique is to compute the reminder of h divided by a carefully chosen prime number.

The described approach is applied both to the songs in the collection and to the queries. With the main goal of efficiency, retrieval is carried out using the *bag of words* paradigm. Similarity between the query and the recordings in the collection is measured by simply counting the number of hashes they have in common. This measure of similarity does not take into account the distribution of hash values. In particular, the occurrence of a chroma hash inside a song and the frequency of a chroma hash across the collection of documents have not been considered. The choice of this particular similarity measure has been motivated by a number of tests using short queries of about 10 seconds, where this simple measure outperformed more complex ones [4].

Since queries of a cover identification task can be complete recordings, the frequency of chroma hashes and their relative position along the song may become a relevant piece of information. In order to handle this issue, long music queries have been divided in overlapping short sub-queries of a fixed duration and for each query an independent retrieval task is carried out. A similar processing is applied to documents, that are divided in overlapping frames with a length comparable to the one of the sub-queries. At the end, the results score of each single retrieval are combined. In particular, preliminary evaluation showed that, in this context, geometric mean outperformed all the other main fusion techniques reported in literature [9].

A problem that may affect retrieval effectiveness is that chroma-based representation is sensible to transpositions. The problem is not dealt with in this paper, as the focus mainly resides in the integration with rhythmic features; it is however part of future work and possible solutions are described in Section IV.

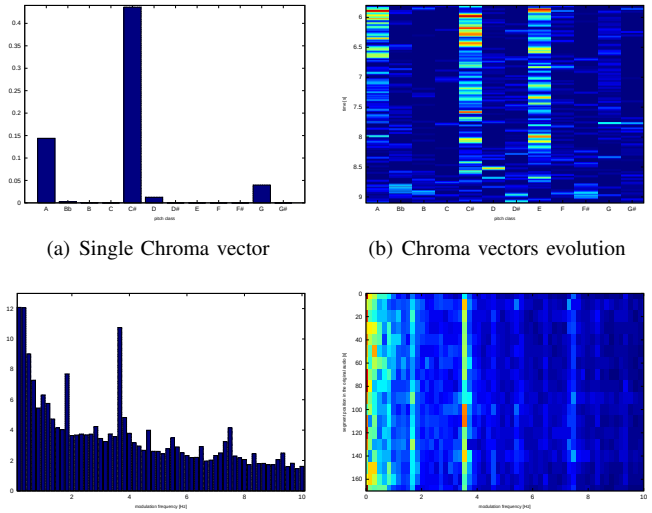
B. Rhythm Histogram features

Rhythm Histogram features [7] are a descriptor for the general rhythmic characteristics of an audio document. In a RH the magnitudes of each modulation frequency bin for all the critical bands of the human auditory range are summed up to form a histogram of "rhythmic energy" per modulation frequency.

In their original form, a single "global" RH represents a whole piece; in our approach, as is the case for Chroma features, a sequence of RHs is computed for each song by segmenting the audio into overlapping windows of 15 seconds, in order to be able to individually match parts of songs which might be characterized by different rhythmic structures (e.g. verse and chorus). Figures 2(c) and 2(d) show the global RH and the sequence of local RHs for David Bowie's "Heroes".

The first step in the computation of the similarity between the songs a and b is the construction of the similarity matrix M, in which each entry m_{ij} is given by the cosine similarity of the i-th RH of a and the j-th RH of b. For each segment of a, the best matching segment of b (that is the one with the highest cosine similarity) is retained, and the mean of these values over all segments of a is computed; a simmetric procedure is then applied to song b and finally the average[1] of these two

[1]A bi-directional comparison procedure is used in order to have a symmetric similarity measure. Experiments however showed that the simpler unidirectional comparison strategy yields similar results.



(a) Single Chroma vector (b) Chroma vectors evolution

(c) Global RH (d) Local RHs

Fig. 2. Chroma and Rhythm Histogram features for D. Bowie's "Heroes"

scores is returned as the similarity of a and b. Experimental results showed that this strategy performs slightly better than the simpler comparison of the global RHs.

It is clear that this approach is computationally intensive, since the cosine similarity of the RHs must be computed for each song in the collection and for each segment pair. Possible optimizations, similar to the ones used for Chroma features, are under investigation. In Section III-C a straightforward strategy for reducing the computational load is proposed, based on the consideration that query songs can be compared to just a small subset of the songs in the collection while retaining the same precision in the results.

C. Feature combination

The similarity score s for a pair of songs, that governs the ranking returned by the system, is computed combining the two scores c and r given by the Chroma features and the Rhythm Histogram features respectively. Two strategies have been used:

- linear combination

$$s = (1-\alpha)c + \alpha r \qquad \alpha \in [0,1] \tag{2}$$

- weighted product

$$s = c^{1-\alpha} r^{\alpha} \qquad \alpha \in [0,1] \tag{3}$$

As pointed out in Section III-C, their performance is similar.

III. Experimental results

Experimental results are presented to show how performances can be improved by combining the scores for the two feature descriptors used. The performances of the system are evaluated using Mean Reciprocal Rank (MRR) as a measure of precision.

A. Test collection

The proposed approach has been evaluated with a test collection of 500 recordings of pop and rock songs, taken from personal collections of the authors. The idea was to have a collection as close as possible to a real scenario. In fact, the indexed documents were all the original version of the music works – i.e., the studio album versions – for which it is expected that metadata are correctly stored and that can be reasonably used in a real system as the reference collection.

The query set included 60 recordings of different versions of a subset of the collection, which were live version by the same artist who recorded the original song and studio or live covers by other artists. We decided to have in the collection one single correct match for each query in order to balance the contribution of each different song. Queries had different durations, generally including the core of the music works – verses and choruses – plus possible introductions and endings, in particular in the live versions. All the audio files were polyphonic recordings with a sampling rate of 44.1 kHz and stored in MP3 format at different bitrates (at most 192 kbps). In fact, in order to simulate a real context, we preferred a compressed format rather than an uncompressed one such as PCM.

B. Individual features results

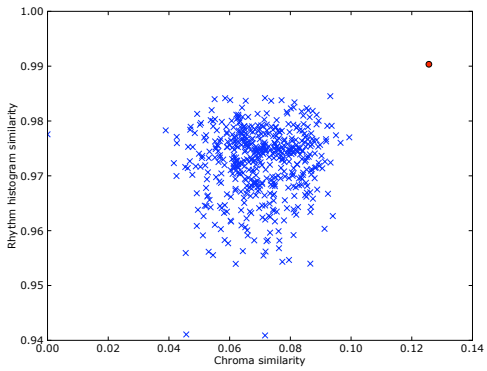
The performance of Chroma features individually is already satisfying, with a MRR of 78.4%. Rhythmic Histogram features on the other hand are less reliable, resulting in a MRR of 34.0%. If the RH features scores are computed directly on the global RH (instead of subdividing the song and computing the best match for each segment) MRR is is 28.5%.

C. Combined features results

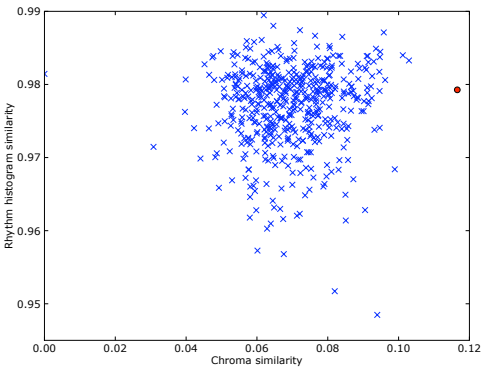
Figure 3 shows typical dispositions of the feature score pairs for some queries; each point in the feature space is associated to a comparison between a query song and the songs in the collection, the red circles being associated to the relevant matches. In particular Figure 3(a) is an example of the best possible situation, in which both Chroma features and Rhythm Histogram features individually rank the correct match for the query song in the first position. Figure 3(b) depicts the most common situation, in which Chroma features correctly identify the match but RH features are misleading; the dual situation is reported in Figure 3(c), which is rare but represents a significant evidence for the usefulness of RH features. Finally Figure 3(d) presents the situation in which neither feature descriptor can correctly identify the best match.

Perhaps the most interesting disposition of score pairs in the feature space is the one depicted in Figure 4: neither feature can identify the matching song by itself, but a combination of the two is indeed able to rank it in the first position.

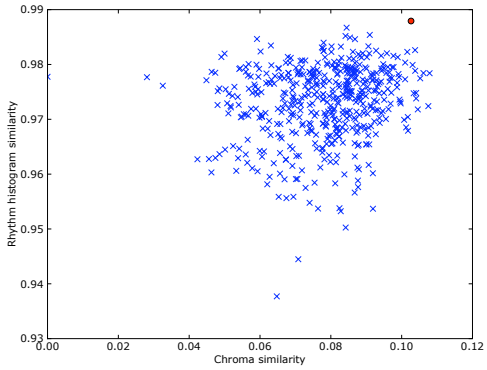
The two approaches to feature combination reported in Equations 2 and 3 have been tested for several values of the parameter α, which weights the influence of RH features in the score, and the resulting MRRs are depicted in Figure 5. For the optimal value of α, the MRR increases from 78.4%



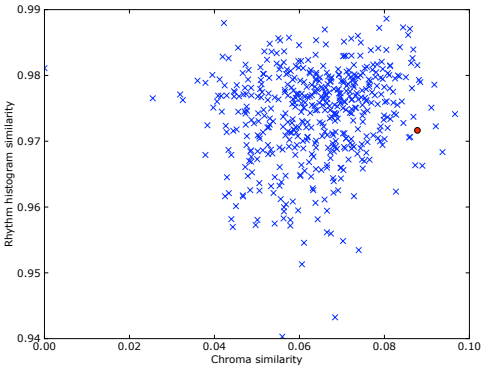
(a) Smoke on the water - Deep Purple



(b) Sweet child of mine - Guns N' Roses



(c) Sweet home Alabama - Lynyrd Skynyrd



(d) You shook me - AC/DC

Fig. 3. Disposition of song similarities in the feature space

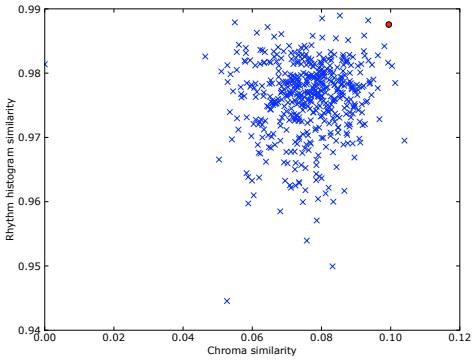


Fig. 4. Disposition of song similarities in the feature space for "All along the watchtower" by Jimi Hendrix

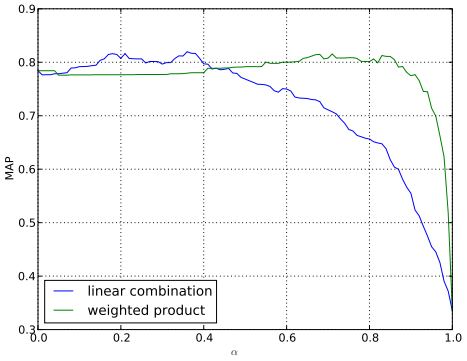


Fig. 5. MRR for the presented approaches to feature combination

(using only Chroma features) to 82.0% and 81.6%, using a linear combination and a weighted product of the features scores respectively. Similar performances are achieved using a single global RH for computing the similarity of songs, with a MRR of 81.5% for the case of the best linear combination of features. Even though the MRR maximum value is located in local peaks of the graphic, which are probably due to the rather small size of the test collection, setting α in a rather large neighbourhood of its optimal value still yields a significant improvement in MRR.

As anticipated in Section II-B, it is clear that performing the comparison of a query song against the whole set of songs in the collection is unfeasible for a large collection,

especially when comparing all the segments against each other. Fortunately Chroma features are able to rank the relevant match in the first positions, and this can be done efficiently thanks to the hashing mechanisms discussed above; an effective solution is to exploit this robustness by reranking only the top t position with the aid of Rhythm Histogram features: with t ranging from 15 to 50 the optimal MRR (82.0%) is unchanged for the collection used. Although the collection is very small, previous experiments with Chroma features on larger collections [4] have shown how the relevant matches for query songs are almost never ranked in very low positions, thus Rhythm Histogram features can be effectively exploited computing them on just a very small fraction of the songs in the collection.

IV. Conclusion

A system for cover identification of pop songs has been presented, focusing on the improvement in identification precision given by the introduction of a rhythmic profile descriptor in addition to the modeling of harmonic content. Many directions for future work are yet to be explored, and the most promising ones are briefly reviewed below.

The modeling of harmonic content still lacks an effective solution for handling the possible transpositions in pitch of cover songs; in fact, if the cover song used as query and the original work stored in the collection are played in different tonalities, they have totally different sets of chroma. This problem can be addressed by considering that a transposition of n semitones will result in a rotation of Chroma vectors of n steps. Then, the tonality issue can be dealt with by simply computing all the twelve transpositions, but at the cost of a loss in precision. Alternatively, a methodology for key finding can be exploited to compute the similarity between the songs in the collection by transposing their chroma features into a reference tonality [3]. The combination of key finding algorithms with the proposed approach will be an important part of future work.

Our investigation of rhythmic content descriptors is still in an early stage. Computational performances are a primary concern, and an efficient implementation of RH retrieval similar to the hash-based implementation of Chroma retrieval is under exhamination.

Finally it is clear how evaluation of the system should be performed on a bigger test collection; this however poses additional issues, not only related to the size of the data that has to be managed, but also to problems regarding music genres, which might have to be dealt specifically: in particular many genres are defined by a very characteristic rhythm (e.g. reggae) thus rhythmic descriptors might be in such cases detrimental to the final performances.

References

[1] P. Cano, M. Koppenberger, and N. Wack, "Content-based music audio recommendation," in *Proceedings of the ACM International Conference on Multimedia*, 2005, pp. 211–212.

[2] F. Kurth and M. Muler, "Efficient index-based audio matching," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 382–395, 2008.

[3] J. Serra, E. Gomez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 6, pp. 1138–1151, 2008.

[4] R. Miotto and N. Orio, "A music identification system based on chroma indexing and statistical modeling," in *Proceedings of International Conference on Music Information Retrieval*, 2008, pp. 301–306.

[5] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *The VLDB Journal*, 1999, pp. 518–529.

[6] M. Slaney and M. Casey, "Locality-sensitive hashing for finding nearest neighbors [lecture notes]," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 128–131, 2008.

[7] T. Lidy and A. Rauber, "Evaluation of feature extractors and psychoacoustic transformations for music genre classification," in *Proceedings of International Conference on Music Information Retrieval*, 2005, pp. 34–41.

[8] D. Ellis and G. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–1429–IV–1432.

[9] E. Fox and J. Shaw, "Combination of multiple searches," in *Proceedings of the Second Text REtrieval Conference (TREC-2)*, 1994, pp. 243–249.

Ensemble of state-of-the-art methods for polyphonic music comparison

David Rizo and José M. Iñesta
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante
Alicante, 03080, Spain
e-mail: {drizo,inesta}@dlsi.ua.es

Kjell Lemström
Dept. of Computer Science
University of Helsinki
FIN-00014 Helsinki, Finland
e-mail: klemstro@cs.helsinki.fi

***Abstract*—Content-based music comparison is a task where no musical similarity measure can perform well in all possible cases. In this paper we will show that a careful combination of different similarity measures in an ensemble measure, will behave more robust than any of the included individual measures when applied as stand-alone measures. For the experiments we have used five state-of-the-art polyphonic similarity measures and three different corpora of polyphonic music.**

I. Introduction

The effectivity of a content-based music comparison or a content-based music retrieval method is mainly down to the appropriateness and degree of success of the underlying similarity measure. To develop an effective musical similarity measure, however, is anything but straight-forward. There is no musical similarity measure that works well in all musical applications. This is partly because musical similarity is, for instance, a culture-dependent, genre-dependent, encoding-dependent, application-dependent — even a user-dependent — issue. Therefore, in literature one can find several musical similarity measures that are effective, in various degrees of success, for different music-related problems.

Typical music-related problems include, for instance, content-based music retrieval, CBMR ("musical pattern matching"); content-based music comparison, CBMC ("musical document similarity") and motive discovery methods ("musical pattern discovery"). One way to make a distinction between methods developed for these problems is to observe their capability to deal either with music containing several *voices* and call such music *polyphonic*, or consider as *polyphonic* music that which simultaneous notes. The most straight-forward methods compare only the melodic lines of *monodies* [1]. One way to deal with polyphonic music is to reduce polyphonic structure in a monophonic form by using a heuristic, such as the skyline algorithm [12]. Naturally, after such a reduction, similarity measures developed for the monodies can be used for this case as well. There are also several methods capable of dealing with polyphony without any reduction, see e.g. [1], [13], [8].

In this paper we deal with symbolically encoded, polyphonic music and focus on the CBMC problem, where the similarity between given two full musical works is to be defined. In order to avoid problems resulting from a use of an unsuitable similarity measure, such as stuck on a local maxima, we suggest an approach that combines several similarity measures. As a careless selection of included similarity measures, such as the one including all possible measures, may result in an excessive running time and a severe bias caused by a "clique of similarity measures", we collect a minimal set of measures having a different aspect on the problem. In a successful compilation, or *ensemble*, of similarity measures, the included individual similarity measures fail and also succeed in different instances of the comparison problem, improving the robustness of the approach.

[1]Monodic compositions either have only a single melodic line, or the composition is dominated by a single melodic line

Fig. 1. Sample score (left) and its skyline reduction (right).

Based on our experiments on three polyphonic music databases and several polyphonic similarity measures reported earlier in the literature, the selection is carried out by using techniques of diversity measurement. To avoid overfitting the resulting ensemble to the data used in the experiments, we have used Kuncheva's overproduce and select method [2].

II. Background

As CBMC has been an active research area for over a decade, in literature one can find several similarity measures developed for the task. In this section we give an overview on similarity measures relevant to our case. The presented methods have been carefully chosen so that they both represent the state-of-the-art and that they give us the needed diversity so that the ensemble to be composed of them would work as well as possible.

We have generated monophonic versions of our three corpora by applying the skyline algorithm (see Fig.1) on the original corpora. In order to boost the diversity among the methods in the resulting ensemble, we have included some monophonic similarity measures that work on the skylined corpora. In addition, the included polyphonic measures are forced to work with both with the skylined and the original versions of the corpora.

Let us now briefly survey the methods and their associated parameters to be adjusted in our experiments. In each of the case, we will also show how the method encodes the short example depicted in Fig.1.

A. Point-pattern / line-segment similarity

Ukkonen et al. [13] considered the CBMR problem and suggested music to be modeled as sets of horizontal line segments in the Euclidean plane, formed by tuples of $\langle$ pitch, duration $\rangle$ (see Fig.2). A reduction of this representation, the point-pattern representation, considers only the starting points

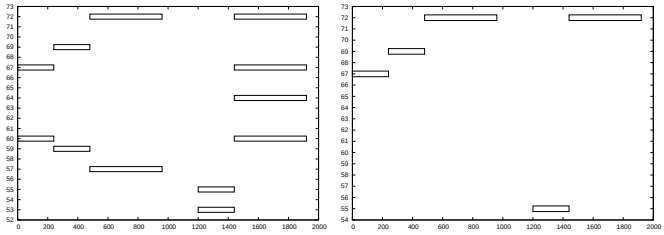


Fig. 2. Line-segment representation (left) and its skyline reduction (right) of Fig.1 on a pitch-against-time Euclidean plane.

Polyphony	r	Sets of [onset, pitch]
Skyline	4	{[0,69], [0,67], [1,72], [2,55], [3,72]}
Skyline	8	{[0,67], [1,69], [2,72], [5,55], [6,72]}
Polyphonic	4	{[0,69], [0,67], [0,60], [0,59], [1,72], [1,57], [2,55], [2,53], [3,72], [3,67], [3,64], [3,60]}
Polyphonic	8	{[0,67], [0,60], [1,69], [1,59], [2,72], [2,57], [5,55], [5,53], [6,72], [6,67], [6,64], [6,60]}

Fig. 3. PROMS representation of Fig.1. Onset is given relative to its position within the bar.

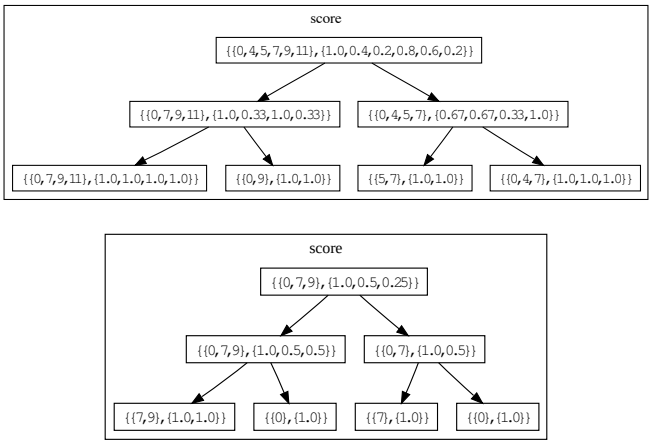


Fig. 4. Tree representations of the polyphonic version (top) and skyline version (bottom) of Fig.1 with $L = 2$. The node labels are of form "{set of pitch classes}, {corresponding cardinalities}".

of the line segments, that is, it is formed of tuples $\langle$ pitch, onset time $\rangle$. Out of their three problems, P2 and P3 are relevant to our case. Intuitively, the similarity of two pieces of music is computed by finding an alignment in the superimposition of two planes representing the considered pieces of music. The alignment should maximize the coincidence rate between the two planes either in point-pattern representation (problem P2) or in line-segment representation (problem P3). Recently, in [3], efficient indexing algorithms for P2 was given. We have included two of them, referred to as P2v5 and P2v6.

B. Bar-specific, quantized point-pattern similarity

Clausen et al. used inverted indices with a representation resembling the above point-pattern representation [1]. The onset times of the notes are quantized to a pre-selected resolution r. Thus, both the pitch and time dimensions become discrete. Moreover, onset times are represented relatively to their metrical position within the musical bar (see Fig.3). The information within a bar constitutes the unit for a query. The retrieval method finds occurrences (total and partial) that have similar metrical positions as the query. Local time and pitch fluctuations cannot be dealt with. Tempo invariance can be obtained by conducting a metrical structure analysis phase and transposition invariance by using a mathematical trick that outperforms the brute-force solution.

In the original algorithm, the approximate search is accomplished by allowing k dissimilarities, in maximum, between the query and the database document. To our needs, where whole pieces of music are to be compared, the original algorithm has been modified to return the normalized number of coincidences in the best alignment. In the sequel, we will refer to this technique as *PROMS similarity*, named after Clausen et al.'s original system.

C. Tree similarity

The tree representation [8] is based on a logarithmic subdivision along the time dimension of a musical bar. Each bar is encoded as a tree and the root-level correspond to the whole note covering the bar. Then, for instance in a binary division, in the next level we have two subtrees corresponding to the two half notes (think of them as fragmentations of the whole note of the root level). This division is continued until the preferred time resolution is met. All trees (i.e. bars) are finally rooted to a common parent representing the whole musical piece. Note that time is implicit in the representation when considering left-to-right ordering of the nodes. The building of a tree is started with pitch information stored in leaves at a level corresponding to its onset (with respect to its position within the bar) and duration. As we are dealing with polyphonic music, the labels of nodes are multisets that contain the cardinality of occurrences of a pitch class in that node. In a second phase, the node labels are bottom-up propagated: The contents of the children are merged in the parent node by using the multiset union operator. Moreover, in order to make the trees smaller, leaves at a level exceeding a depth threshold L are pruned.

The similarity of two pieces of music represented in this way is measured by combining Selkow's tree-edit-distance [10], that accounts for structural similarity, with a multiset distance measure accounting for musical surface level similarity. In [9] several multiset similarity measures to this end have been suggested.

D. n-gram similarity

Linear strings have often been used for encoding monophonic music. In doing so, one may also harness the general string matching techniques for the problem. In [11], Uitdenbogerd suggested the use of n-gramming on strings over a pitch interval alphabet, where the original interval values were reduced by performing $\mod 12$ over them.

We have implemented four methods based on the string matching framework performing over the aforementioned interval alphabet. The first uses the classical Levenshtein distance [4] (here called *NGEdit*). Then we have two modifications of the Levenshtein distance, one using a simple local alignment (called *NGLocal*) and one using so-called start-match alignment (called *NGMatch*). The last one is Uitdenbogerd's n-gramming technique.

Representation	Actual value
String	ophr
2-grams	{hr, op, ph}
3-grams	{oph, phr}

Fig. 5. A string and the corresponding 2- and 3-grams of skyline score in Fig.1

E. Graph similarity

In [7], Pinto modeled monodies by using directed graphs over the pitch profile. In his graph, each pitch class is associated with a node, and each melody transition from a pitch class to another with a label. The label represents the frequency of the interval, between the two associated nodes, within the represented monody (see Fig.6). The graph represents also the interval from the last note of the melody to the first one.

The similarity of the graphs is measured by using a their laplacian spectra as a feature vector. Laplacian spectra is invariant under shifts of the columns, that is, invariant under musical transpositions.

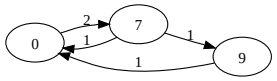


Fig. 6. Graph representation of skyline score in Fig.1

III. Classifier combination

As argued above, any single musical representation or similarity algorithm cannot give succesful results in all our cover version identification tasks. In [2], Kuncheva showed that a carefully chosen combination of classifiers, compiled into so-called *ensembles*, will result in more robust systems that perform at least at a comparable level when compared with any individual classifier. There are some constraints, however, that limit the performance of such ensembles, the most important being that the classifiers of the ensemble should differ in the kind of errors they make. This led us in chosing very different approaches, the ones described in Section II.

To measure the diversity amongst the chosen classifiers in the ensemble, we have used the *interrater-agreement* κ: the lower the κ, the higher the disagreement and hence higher diversity (see eq.(10.8) in [2]). The measure is computed for each pair of classifiers, and it is inversely proportional to the diversity.

To choose the ensemble from a set of available classifiers, Kuncheva proposes to use the *overproduce and select* method. In our case this works as follws: Given a training set and the classifications of all classifiers, choose the ones giving the highest diversity and the lowest average error rate. We obtain such a set of classifiers by computing a *pareto-optimal* set. The latter set is computed as follows: the κ and average error rate for each pair of classifiers is computed, then only those pairs with both best κ and average error rate are kept (they are said to be *non-dominated*). The *pareto-optimal* set is composed by the classifiers that form that kept pairs.

Once the most suitable classifiers are selected, any of the combination scheme in [5] can be used. In this paper, we have used the raw voting scheme.

IV. Experiments

The experiments are designed to check the suitability of the combination of different polyphonic music similarity paradigms in comparison to individual methods. To this end, first the performance of each individual method with all its possible setups has been tested, then the diversity of all possible pairs of classifiers has been studied using the κ statistic. Given this diversity measure the best classifiers have been chosen using several selection schemes, and combined using a voting ensemble.

A. Corpora

Three different corpora of cover versions have been collected in order to show the behaviour of the methods with different data. Each corpus is organized into songs or classes. For each song there is a main prototype and a set of variations or cover versions.

The first corpus, called *ICPS*, has 68 MIDI files corresponding to covers of the incipits of seven musical works: Schubert's "Ave Maria", Ravel's "Bolero", the children songs "Alouette", "Happy Birthday" and "Frère Jacques", the carol "Jingle Bells" and the jazz standard "When The Saints Go Marching In". All the works in this corpus have a similar kind of accompaniment tracks with a prominent melody track.

The second corpus, named *VAR*, consists of 78 classical works representing variations of 17 different themes as written by the original composer: Tchaikovsky "variations on a rococo theme" op.33, Bach "English suites" BWV 806-808 (suite 1 courante II, suite 2, 3, and 6 sarabande), and Bach "Goldberg variations". In this case, the variations are founded mostly on the harmonic structure of a main theme.

The third one, called *INET* is made of 101 whole MIDI files downloaded from the Internet corresponding to 31 different popular songs. It contains real-time sequences with mistakes and different arrangements of the original versions.

B. Quality measurement

The experiments have been performed using a query / candidates scheme, i.e., given a corpus, for each song its main prototype (acting as query) is compared with all the cover versions of all songs. The similarity values of all the comparisons are ordered and following a 1-NN rule, the best value is taken as the answer. This answer is correct if it corresponds to the same song of the query. Any other situation is considered as an error.

Thus, the success rate is measured as the rate of correct answers for all queries or main prototypes in a corpus.

C. Individual performances of methods

In the following points, the results for all possible setups of the tested methods are plotted. The reader can remind the meaning of each parameter in the figures in the introduction of the methods in Section II above.

The monophonic methods (graph and n-grams) have been fed with a skyline reduction of the corpora. The polyphonic methods have been tested using both the original corpora and also the same skyline reduction. In the plots, the corpora with the skyline applied are denoted with a "M-" prefix.

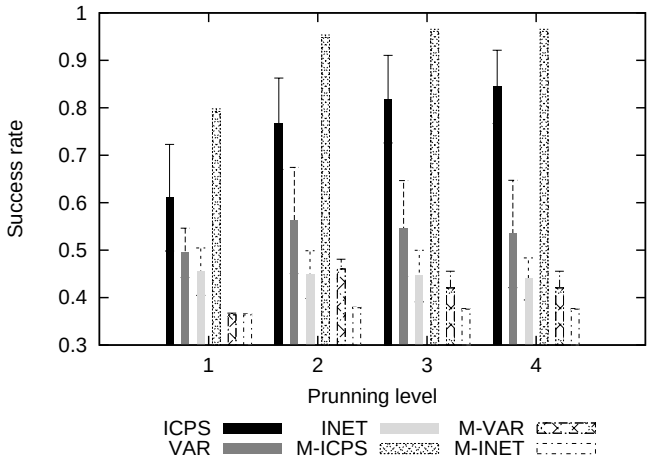


Fig. 7. Prunning levels

1) Trees results: The results in Fig. 7 show the averages and standard deviations for each pruning level with all multiset distances used, resulting in 44 different classifiers grouped into those 3 pruning levels. It can be noticed that for the corpus with a more evident melody line, the *ICPS*, the skyline reduction improves the classification rate. This is not the case for the other corpora, where a more important harmonic component is found, and the polyphony is required to identify versions. For the plots, the best pruning level after propagation is $L = 2$. For monophonic corpora, the system may require larger trees with the increase of classification times.

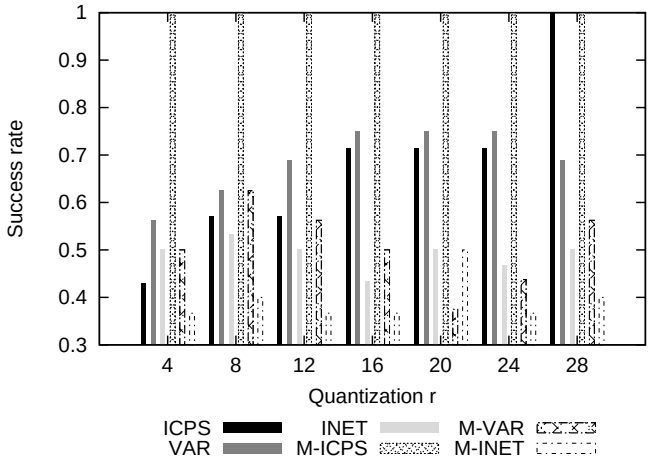


Fig. 8. PROMS, the x axis represents the different resolutions r in which each song bar is quantized

2) PROMS results: The different setups of the PROMS depend on the quantization per bar. We have tested from 4 (which means a resolution of quarter note for a 4/4 meter) to 28 where the system seems to stabilize.

The PROMS representation is very sensitive to quantization. This quantization produces chords with closer notes, being this fact noticeable in the behaviour of the monophonic corpus in front of the polyphonic one (see Fig. 8): a quantization of polyphonic content leads to too dense chords, and as the r raises, it gets less dense. On the other hand, more quantization of monophonic content helps in the classification.

Anyway, it seems that the best average setup is that of r=28, that is not tight to any meter as r=12 is to 3/4, or $r = 16$ to 4/4.

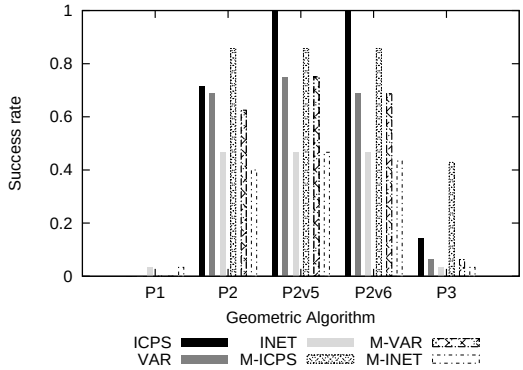


Fig. 9. Geometric algorithms

3) Geometric results: The five different geometric algorithms results are plotted in Fig. 9, showing that the best algorithms are P2v5 and P2v6. It is remarkable the robustness of these classifiers against the polyphony, behaving comparably with the original content and the "skylined" one.

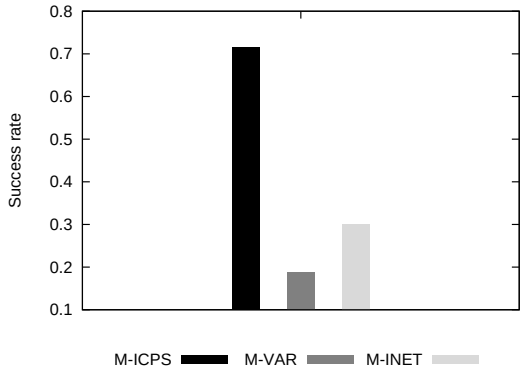


Fig. 10. Graphs

4) Graphs results: This method works only with the pitch component of monodies, so it uses limited information. Results shown in Fig. 10 evidence this fact. The only well managed corpus is that with a clear melody, *ICPS*. However, the best quality of this paradigm is its performance time, so it is worth to include it in the ensembles.

5) n-grams results: The results of the string matching approach (Fig. 11) and the n-grams classifying (Fig. 12) show that the later can deal better with skylined content than a simple string matching approach. The best results are located around the 5-grams. Anyway, as was previously introduced in [6], the best option is to combine various n-gram sizes, what will be done in the ensembles. Being the string methods very fast, we will include them in the combinations.

6) Best of each method and each corpus: In Table I the best of each method setup is detailed for each corpus. None

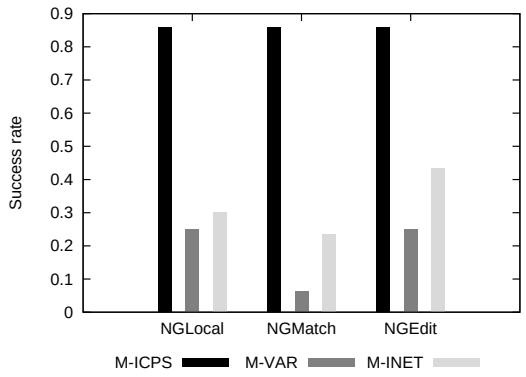


Fig. 11. Uitdenboderg Strings

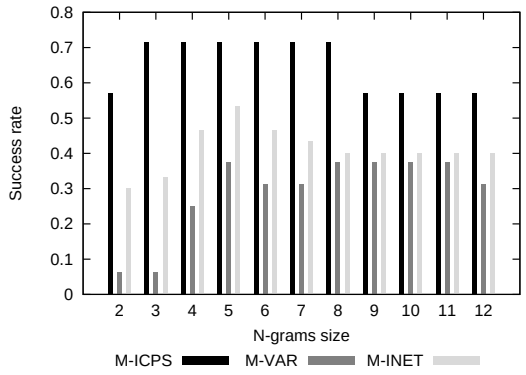


Fig. 12. Uitdenboderg ngrams

of the paradigms can be stated as the best for all situations, so a combination that takes advantage of the qualities of each seems to be the best option.

D. Ensemble methods

In this experiment, the ability of ensemble methods to improve overall classification rates has been tested. In order to choose only the best combinations in terms of number of classifiers included some selection methods have been compared.

These selections are based in choosing either the most diverse M classifiers, i.e., those giving the most different classifications, or choosing the classifiers with best trade-off between diversity and average error rate. Both selections are based on a plot of the κ statistic. Figures 13, 14, and 15 represent, for each corpus, that κ vs. average error rate for each pair of classifiers. The most diverse M classifier selection chooses the M rightmost points in these plots. The *pareto-optimal set* (aka. *Par.*) is shown by square points in the plots.

Fig.16 shows the behaviour of selection methods with the three corpora. From the plot, it can be stated that the best option is the use of the *pareto-optimal-set*, for the success rates and the number of classifiers involved.

Finally, the results of this ensemble compared to the best of individual methods (Table II) show that the most robust method is the ensemble, being significant the improvement in the most difficult method, the *INET*.

TABLE I
BEST SUCCESS RATES

Corpus	Method	Best method setting	Result (%)
VAR	Geometric	P2v5	75
M-VAR	Geometric	P2v5	75
M-VAR	Graph		19
VAR	PROMS	$r = 24$	75
M-VAR	PROMS	$r = 8$	63
VAR	Trees	Cosine Dist, $L = 2$	75
M-VAR	Trees	Cosine Dist, $L = 2$	75
M-VAR	String matching	Local, Edit	25
M-VAR	n-grams	$n \in \{5, 8, 9, 10, 11\}$	38
ICPS	Geometric	P2v5 and P2v6	100
M-ICPS	Geometric	P2,P2v5,P2v6	86
M-ICPS	Graph		71
ICPS	PROMS	$r = 28$	100
M-ICPS	PROMS	Any r	100
ICPS	Trees	Log dist, $L = 1$	100
M-ICPS	Trees	Var. dist, $L = 4$	100
M-ICPS	String matching	Any	86
M-ICPS	n-grams	$n \in [3..8]$	78
INET	Geometric	P2v6	47
M-INET	Geometric	P2,P2v5,P2v6	47
M-INET	Graph		30
INET	PROMS	$r = 8$	53
M-INET	PROMS	$r = 20$	50
INET	Trees	Log dist, $L = 1$	53
M-INET	Trees	Harmonic mean, $L = 4$	43
M-INET	String matching	Edit	43
M-INET	n-grams	n=5	53

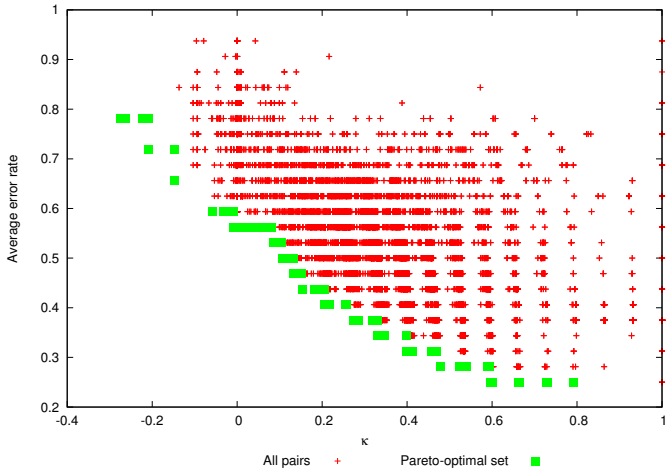


Fig. 13. κ vs. average error rate in corpus *VAR*

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we have considered five different approaches to measure musical similarity. In order to study how they perform, both as a stand-alone measure and as part of an ensemble measure, we have experimented on them using three distinct music corpora. An ensemble measure, or classifier, produced by *overproduce and select* method, has been shown to be superior to any of the individual stand-alone classifiers.

Now that we know the ensemble classifier to give good results, our next step is to build a large corpus with a good variety in genres. Having such a corpus, we would be able to apply standard cross-validation rules in training and testing with new data. We will also apply some more sophisticated combination schemes that are expected to improve the rates of correct classification.

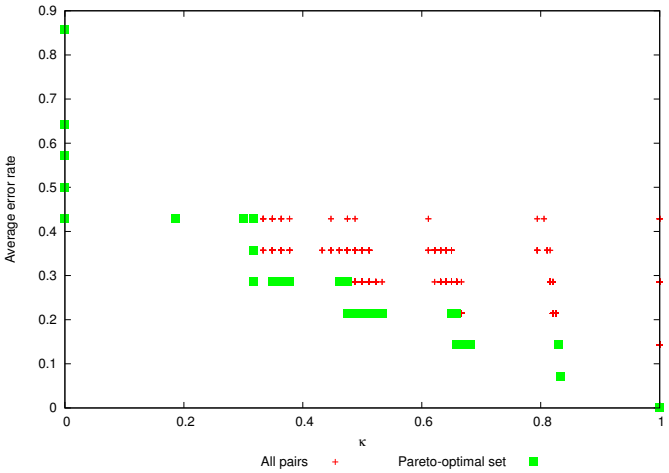


Fig. 14. κ vs. average error rate in corpus *ICPS*

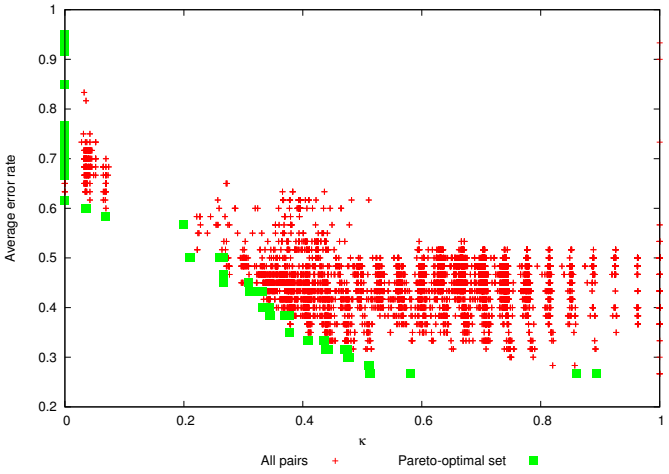


Fig. 15. κ vs. average error rate in corpus *INET*

TABLE II
BEST SUCCESS RATES OF ALL INDIVIDUAL AND COMBINED METHODS

Corpus	Method	Best Result (%)
VAR	Individual	75
VAR	Combined	84
ICPS	Individual	100
ICPS	Combined	100
INET	Individual	53
INET	Combined	82

VI. ACKNOWLEDGMENTS

Kjell Lemström was supported by the Academy of Finland (Grant 108547).

David Rizo and José M. Iñesta are supported the Spanish Ministry projects: TIN2006–14932–C02 and Consolider Ingenio 2010 (MIPRCV, CSD2007-00018), both partially supported by EU ERDF. Special thanks to José F. Bernabeu for his technical support, Iman Suyoto, and Alberto Pinto for their collaboration.

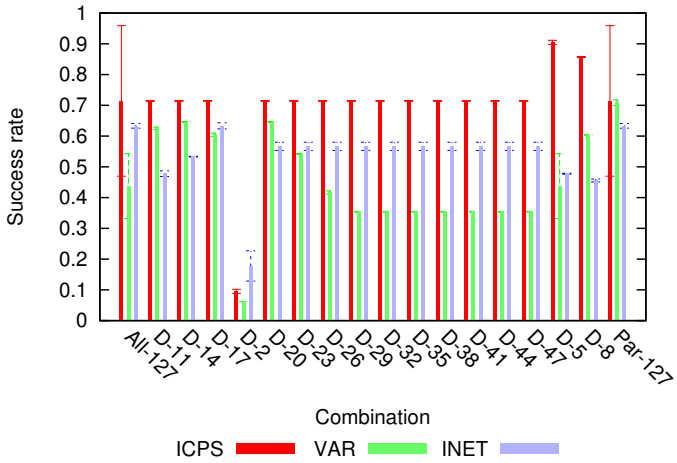


Fig. 16. Classifier selection methods. *All* means no selection, that is, use all classifiers. DM is the ensemble built from the classifiers included in the M most diverse pairs. *Par.* is the *pareto-optimal-set*.

REFERENCES

[1] Michael Clausen, Ronald Engelbrecht, D. Meyer, and J. Schmitz. Proms: A web-based tool for searching in polyphonic music. In *ISMIR*, 2000.

[2] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, July 2004.

[3] Kjell Lemström, Niko Mikkilä, and Veli Mäkinen. Fast index based filters for music retrieval. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, September 2008.

[4] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[5] F. Moreno-Seco, José M. Iñesta, P. Ponce de León, and L. Micó. Comparison of classifier fusion methods for classification in pattern recognition tasks. *Lecture Notes in Computer Science*, 4109:705–713, 2006.

[6] Nicola Orio. Combining multilevel and multifeature representation to compute melodic similarity. MIREX, 2005.

[7] Alberto Pinto, Reinier H. van Leuken, M. Fatih Demirci, Frans Wiering, and Remco C. Veltkamp. Indexing music collections through graph spectra. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*, Vienna, Austria, September 2007.

[8] David Rizo, Kjell Lemström, and José M. Iñesta. Tree structured and combined methods for comparing metered polyphonic music. In *Proc. Computer Music Modeling and Retrieval 2008 (CMMR'08)*, pages 263–278, Copenhagen, Denmark, Copenhagen, Denmark, May 19-23 2008.

[9] David Rizo, Kjell Lemström, and José M. Iñesta. Tree representation in combined polyphonic music comparison. *Lecture Notes in Computer Science, selected papers from the CMMR 2008 (to appear)*, 2009.

[10] Stanley M. Selkow. The tree-to-tree editing problem. *Inf. Process. Lett.*, 6(6):184–186, 1977.

[11] Alexandra L. Uitdenbogerd. N-gram pattern matching and dynamic programming for symbolic melody search. MIREX, 2007.

[12] Alexandra L. Uitdenbogerd and Justin Zobel. Manipulation of music for melody matching. In *ACM Multimedia*, pages 235–240, 1998.

[13] Esko Ukkonen, Kjell Lemström, and Veli Mäkinen. Sweepline the music! In *Computer Science in Perspective*, pages 330–342, 2003.

Measuring Harmonic Similarity Using PPM-based Compression Distance

Teppo E. Ahonen
Department of Computer Science
University of Helsinki
Helsinki, Finland
Email: teahonen@cs.helsinki.fi

***Abstract*—Normalized compression distance (NCD) is a metric where the similarity between two objects is calculated based on their compressibility with a lossless compression algorithm. Here, NCD is used to measure the similarity between chord sequences estimated from audio files using PPMZ as compression algorithm. Since NCD is normalized and captures the dominant similarity of the objects, we assume that it is not hindered by the differences in structures and lengths of the different versions of musical documents. The method is evaluated in cover song identification, and results suggest that the approach has potential for this task, though the shortcomings of chord estimation have a distinct effect on the quality of identification.**

***Index Terms*—normalized compression distance, harmonic similarity, cover song identification, PPM**

I. Introduction

In recent years, the rapidly growing amount of digital audio libraries has increased researchers' interest towards retrieving music based on its content rather than its metadata. Different methods for measuring similarity based on the content of the musical documents have been developed by various researchers in music information retrieval (MIR) community. Still, the question of measuring essential similarity remains unsolved.

Cover song identification is a music information retrieval task where the goal is to determine whether two pieces of music are different interpretations of a same composition. Successful cover song identification yields important information on musical similarity, as human ear can easily recognize even remarkably diverse interpretations of a known composition. Since cover versions especially in popular music are often intentionally different from the original versions, the task is a very difficult challenge for computers. Measuring similarity must be based on so-called mid-level tonal features, instead of low-level audio signal features.

Normalized compression distance (NCD) [1] has been successfully used for measuring similarity in symbolic music data [2]. So far there has been very little research where NCD is used for features derived from audio data. Here, we present a method that uses NCD to measure similarity between chord sequence approximations.

The rest of this paper is organized as follows. First, in Section II, we give a brief tutorial on NCD. Then, in Section III we describe our NCD-utilizing approach to cover song identification. In Section IV we present and analyze the results of our evaluation tests. Finally, we present conclusions in Section V.

II. Normalized Compression Distance

We do not go into the details of NCD in this paper. For those interested on the information theoretic background of NCD and more accurate definitions of properties required from compression algorithms, we suggest to read the original articles [1], [3]. Here, we present only the intuitive basis of the metric.

In [1], a similarity measure called normalized information distance (NID) was presented. NID uses Kolmogorov complexity to measure the distance between objects. Kolmogorov complexity of an object is the length of the shortest binary program that outputs the object on a universal computer.

Denote $K(x)$ as the Kolmogorov complexity of string x, and $K(x|y)$ as the conditional Kolmogorov complexity of string x given string y as input. NID for strings x and y is denoted as [1]

$$NID(x,y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}}. \quad (1)$$

The NID is a universal similarity metric, since it minorizes every computable metric [1]. However, as Kolmogorov complexity is non-computable, NID can not be calculated.

Denote $C(x)$ as the length of a string x when compressed using a lossless data compressor C. Also denote $C(x|y)$, the conditional compressed information in x that can be defined as $C(x|y) = C(xy) - C(y)$, meaning the amount of bits of information in x related to y. We can approximate Kolmogorov complexity K with C: the better we are able to compress, the closer $C(x)$ is to the Kolmogorov complexity of x. Now, we can approximate Equation (1) and normalized compression distance for strings x and y can be defined as [1]

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}. \quad (2)$$

The NCD is a metric, but not universal; it can still be called *quasi-universal* since it minorizes every computable distance up to an error depending on the quality of the compressor to approximate Kolmogorov complexity [3]. Because NCD uses only standard compression algorithms, no a priori information of the data is needed and NCD can be used for any domain.

Thus, NCD has been successfully used for several clustering tasks in different domains varying from genome data to optical character recognition [3].

III. System for Cover Song Identification

We have designed and implemented a system for cover song identification that uses NCD to calculate the similarity between chord sequences estimated from audio files.

We presented an early version of our system in [4]. Since then, we have made several upgrades that have had a positive effect on the identification results. Most importantly, we have included a PPM compressor in addition to gzip and bzip2. Since PPM is generally a more effective compressor, it yields a better approximation of Kolmogorov complexity.

Apart from this, we have also tested different encodings to represent chord sequences. This is important, as the chord sequence should express the essential information in tonal content of the piece of music, and each encoding captures slightly different aspects of music.

A. System description

Our system takes two lists of audio files as input, one consisting of m target pieces of music and another consisting of n query pieces of music. As a result, the system returns an $n \times m$ distance matrix. A blueprint of the components of the system is presented in Figure 1.

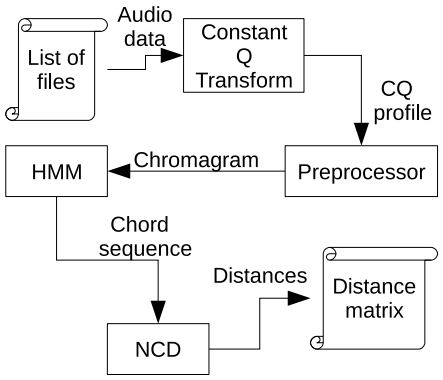


Fig. 1. Blueprint of the system components

Each audio file is converted into a chromagram, a sequence of 12-position vectors that indicate the spectral energy distribution of pitch classes. Chromagram, or pitch class profile, can be obtained by using constant Q transform [5]. We use window length of 4096 samples and hop size of 1024 samples, limit the frequencies between four octaves starting from 65.4 Hz and use 36-bin resolution in an octave. In the preprocessing phase, the constant Q profile is turned into a chromagram by folding all octaves together and normalizing each vector.

Several chord estimation algorithms use beat estimation to reduce mislabelings caused by transients. In our tests, however, using frame-based chord sequences led into better results. This could be due to two reasons. Firstly, faulty beat estimation can cause even more noise in the estimated chord sequences. Similar observation was made in [6], where tactus-based chord estimation weakened the results in cover song identification. Secondly, tactus-based sequences are shorter. Keeping in mind that the error between $C(x)$ and $K(x)$ minimizes as file length increases [3], longer sequences lead into better estimation of NID.

The chroma vectors are given as observations to a 24-state ergodic hidden Markov model (HMM), with initial parameters set as suggested in [7]. Each state of the model represents a major or minor triad chord, with one state for each root semitone and transition probabilities between states are initially set according to a double-nested circle of fifths. The model is trained with the EM algorithm and the most likely path through the states is obtained with the Viterbi algorithm.

Estimated chord sequences are written into text files. We have experimented several different ways to represent the chord sequences. Easiest way to represent a chord sequence would be to represent each chord with a distinctive character. However, cover versions are occasionally in different key, making such representation unsuitable. This could be solved by estimating the key of the piece of music and transposing every piece of music into common key, but as with the beat estimation, we have empirically discovered that unsuccessful key estimation will have a negative impact on the results. A brute force approach would be to calculate distance between the target and all 12 transpositions of the query sequence and select the one with the shortest distance, but we consider this approach to be too time-consuming and cumbersome. Thus, we chose to use a representation that is key invariant.

Previously, we discovered empirically that the most suitable representation for chord sequences was a representation where the change between two subsequent chords is denoted as three characters: first two describing the semitone difference between chords' root notes and the third describing if there is a change between major and minor chord or not. Later, we shortened this even more by denoting each chord change by one character. As the difference between two chords' root notes is between $[-5, +6]$ and there either is a change between major and minor chord or not, the size of the alphabet for our sequences is 24.

Finally, the NCD values of Equation (2) are calculated for each pair of written chord sequences and a distance matrix is constructed from these.

B. PPM

PPM (Prediction by Partial Matching) [8] is a data compression algorithm scheme that uses context modeling for predicting the next symbol to be encoded. There are several variants of the scheme, our choice is PPMZ.

Since PPM is a statistical compressor, it is more robust and thus more suitable for NCD than commonly used dictionary or block-sorting compressors such as gzip or bzip2. Similar observations have been made in [9]. In their benchmark tests for compressor idempotency they noticed that when using PPMZ, the sizes of objects do not affect the distance and the distance is also in general very small [9].

We used the pzip algorithm[1] as our PPMZ implementation. PPMZ is clearly slower than gzip or bzip2, but achieves more efficient compression and thus is a better estimation of Kolmogorov complexity, which makes up for the time consumption.

IV. Evaluation

The system was tested on genuine audio data. We used the "covers80" dataset by Daniel Ellis[2]. This is a 160-file collection of 80 pairs of original pieces of music and their cover versions.

We use the former half of the 160 pieces of music as queries and the latter half as as a target dataset. As the dataset is rather small, we only considered a result set of size three.

A. Results

Our results for 80 queries with a result set of three are given in Table I. We measured the number of queries where the distance between the original version and the corresponding cover version was the smallest (Match), the number of cases where the cover version was included in the result set (Top-3) and the mean of average precisions (MAP).

TABLE I
Results for covers80 dataset with a result set of size three

	Range	PPMZ	bzip2	gzip
Match	[0-80]	16	14	14
Top-3	[0-80]	28	26	23
MAP	[0-1]	0.269	0.231	0.227

Based on these results, it is clear that efficient compression benefits the identification as the PPMZ compressor is superior to gzip and bzip2. Still, the results leave room for analysis and possible further development.

B. Golden standard

To evaluate the aptitude of NCD for cover song identification we annotated manually a subset of the dataset to see how well NCD works when data is not distorted. We took the chord sequences provided by the HMM and corrected all mislabeled chords before writing the sequences into files.

Hand-writing correct sequences is not only laborious, but also rather difficult as there rarely exist ground truth chord sequences for cover versions. Thus, we annotated only 12 pairs of pieces of music for our tests. Though the subset is rather small, we believe it is sufficient to demonstrate that NCD can well be used for identification tasks.

The distances between original and cover versions with both annotated and genuine chord sequences are depicted in Figure 2. It is evident that the distances are in most cases notably larger with the estimated sequences.

The test proved that NCD can successfully identify cover versions when chord sequences are successfully estimated and contain very little noise: the distances between the original and cover version were smallest in all 12 cases. It should be noted, that the pieces of music had similar chord progressions, but the musical structures and the lengths of the sequences were different, which was not an obstacle in identification. We also tested another representation: we transposed the chord sequences into a common key of C major, and wrote the sequences into files using notation where each chord was represented as a single character. Results were similar, thus implying that this representation would also be usable, if key estimation would be reliable.

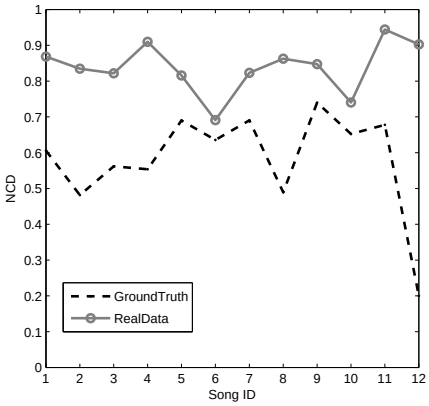


Fig. 2. Distances between 12 original and cover version pairs of annotated and actual chord sequences

However, in real life the chord sequences are always only approximations with more or less errors: mislabeled chords and distortion caused by non-tonal instruments such as drums in popular music. This causes the chord approximations to be noisy, which in turn weakens the Kolmogorov complexity approximation of the compressor, as the more noisy the data is, the less it can be compressed. In [10] it was shown that NCD is robust against noise, but here the compressed sequences are short and constructed from a small alphabet, thus enabling even small amount of noise to have a major impact on the identification. When testing the same 12 musical document pair set with actually estimated chord sequences, we were able to identify the cover version in eight cases, implying the effect the noise can have in even a small set of data.

Also, as PPMZ is a statistical compressor, it was more tolerable to noise in sequence approximations than gzip or bzip2. Differences between the average distances between correct musical document pairs with different algorithms are presented in Table II.

TABLE II
Average NCD values of correct version pairs for annotated ground truth files and actual estimated chord sequences

	PPMZ	bzip2	gzip
Ground truth	0.5815	0.5960	0.6477
Real data	0.8385	0.8724	0.8828

[1]http://muq.org/ cynbe/compression/ppm.html

[2]http://labrosa.ee.columbia.edu/projects/coversongs/covers80/

C. Testing with synthetic audio

In order to evaluate the performance of the system with cleaner data, we obtained MIDI versions of 30 original pieces of music of the dataset and transformed the MIDI files into audio using Timidity++. The MIDI-based musical documents were far less arranged and orchestrated, thus containing far less elements that hinder the HMM chord estimation and in turn resulting in more accurate chord sequences, situated somewhere between the manually annotated and the authentic audio-based estimations.

We used the original audio versions as the dataset and the MIDI versions as queries. Also, we ran a test using MIDI versions as queries and cover versions of the audio versions as the dataset. To compare this performance with purely audio-based identification, we ran third test using cover versions as queries and original versions as the target dataset.

Results for these tests are presented in Table III. We measured the number of correct identifications (Match) and the average distance between correct version pairs (Distance).

TABLE III
RESULTS FOR 30 MUSICAL DOCUMENT PAIR QUERIES WITH MIDI-BASED AUDIO USING PPMZ

Query	Target	Match	Distance
MIDI	Originals	12	0.8528
MIDI	Covers	8	0.8537
Originals	Covers	5	0.8905

Results imply that NCD does perform somewhat better with the MIDI-based sequences, as these are less flawed. However, when both query and target data suffer from severe noise, results are distorted and the performance weakens significantly.

V. CONCLUSIONS

We have presented a method for calculating similarity between musical documents using normalized compression distance for chord sequences approximated from audio data. Results imply that the approach has potential, but the distortion on the approximated chord sequences has a negative effect on identification.

When testing with manually labeled ground truth chord sequences, NCD was able to successfully identify cover versions from a small set of musical document pairs. However, results were not as strong with authentic chord approximations, which implies that the weakest link of the system is the chord estimation. Similar observations were made when using audio files created synthetically from MIDI as queries, as these chord sequences contained less noise.

It is clear that a more precise chord estimation would have a positive impact on our approach. Currently, the overall noise on chord sequences weakens the compression, which in turn weakens the distance calculation. Also, successful key estimation could possibly benefit identification, as sequences could be represented in a simple, distinctive way.

Another approach might be to focus on the chroma vectors themselves instead of estimating the chord sequences from them. This is a common approach on some of the current state-of-the-art cover version identification systems, such as [11]. Relying only on chord sequences is probably too limiting, as similar chord sequences can be found in various different pieces of music that still are otherwise unique compositions. Also, many pieces of music consist of chords such as augmented, diminished and suspended chords, which are not managed very well with a 24-state HMM. As chromagram captures not only harmonic but also melodic information, it represents the mid-level features of a piece of music better than a chord sequence. Thus, quantizing chroma vectors without attempting to estimate the chord sequence could provide a suitable representation for NCD-based identification.

In future, we will adopt NCD for several other MIR tasks that require measuring similarity between musical features.

ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland (Grant #129909).

REFERENCES

[1] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi, "The similarity metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, December 2004.

[2] R. Cilibrasi, P. M. B. Vitányi, and R. de Wolf, "Algorithmic clustering of music based on string compression," *Computer Music Journal*, vol. 28, no. 4, pp. 49–67, December 2004.

[3] R. Cilibrasi and P. Vitányi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1523–1545, April 2005.

[4] T. E. Ahonen and K. Lemström, "Identifying cover songs using normalized compression distance," in *Proceedings of the International Workshop on Machine Learning and Music*, Helsinki, Finland, July 2008.

[5] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of Acoustic Society of America*, vol. 89, no. 1, pp. 425–434, January 1991.

[6] J. P. Bello, "Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats," in *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, September 2007, pp. 239–244.

[7] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *Proceedings of 6th International Conference on Music Information Retrieval*, London, UK, September 2005, pp. 304–311.

[8] J. G. Cleary and I. H. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Transactions on Communications*, vol. 32, no. 4, pp. 396–402, April 1984.

[9] M. Cebrián, M. Alfonseca, and A. Ortega, "Common pitfalls using normalized compression distance: what to watch out for in a compressor," *Communications in Information and Systems*, vol. 5, no. 4, pp. 367–384, 2005.

[10] ——, "The normalized compression distance is resistant to noise," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1895–1900, May 2007.

[11] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1138–1151, 2008.

The NEUMA Project: Towards Cooperative On-line Music Score Libraries

L. Abrouk[1], H. Audéon[2], N. Cullot[1], C. Davy-Rigaux[2], Z. Faget[3,4], D. Gross-Amblard[1], H. Lee[1],
P. Rigaux[3], A. Tacaille[5], E. Gavignet[1], V. Thion-Goasdoué[3]
[1]LE2I, Univ. Dijon, Fr. [2]IRPMF, CNRS/BnF, Paris, Fr. [3]LAMSADE, Univ. Paris-Dauphine, Fr. [4]ARMADILLO, Fr. [5]PLM, Univ. Sorbonne, Fr.

***Abstract*—The NEUMA project (*http://neuma.irpmf-cnrs.fr*) aims at designing and evaluating an open cooperative system for musician communities, enabling new search and analysis tools for symbolic musical content sharing and dissemination. The project is organized around the French CNRS laboratory of the *Bibliothèque Nationale de France* which provides sample collections, user requirements and expert validation. The paper presents the project goals, its achitecture and current state of development. We illustrate our approach with an on-line publication of monodic collections centered on XVIIe century French liturgic chants.**

I. INTRODUCTION

Current musical search engines mostly rely on audio files. Extracting significant information from such files to produce some structured description is a difficult task. Another approach is to exploit *symbolic* music representation, usually derived from musical scores. This representation contains a detailed description of all the information required to produce a precise rendering. Of course, interpreting this information is largely a matter of human expertise, and a software can hardly manipulate anything but a minor part of the intentional content conveyed by the symbolic notation. However, such manipulations can obviously go far beyond the mere on-demand printout of score copies. They include structure-based navigation, melodic fragments identification and retrieval, support for musical analysis, large-scale handling of scores databases, and community sharing and dissemination. Exploring the precise limits on such automatic or semi-automatic functionalities constitutes an exciting research area.

The goal of the NEUMA project (*http://neuma.irpmf-cnrs.fr*) is to address some of the main issues raised by the constitution of on-line digital scores libraries, and in particular:

1) create and maintain repositories of large-scale musical collections, along with appropriate browsing, searching and rendering functionalities;
2) explore automatic and semi-automatic music analysis tools applied to symbolic notation;
3) investigate problems related to copyright protection;
4) and, finally, experiment web-based sharing of musical scores archives, including cooperative development, data reconciliation and annotation.

NEUMA is a three-years project founded by the French *Agence Nationale de la Recherche* (ANR), and is built around the IRPMF laboratory (*http://www.irpmf.cnrs.fr*), a CNRS/BnF research institute devoted to the preservation and dissemination of Western music heritage. The project started in january 2009 and is currently in its early stages of development.

In the present paper, we first explain the main goals of NEUMA (Section II). We then outline the architecture of the project in Section III. Section IV is devoted to the first corpuses published by NEUMA. Finally, Section V discusses related works and Section VI provides the project's roadmap.

II. PROJECT GOALS AND REQUIREMENTS

NEUMA intends to support the development of digital collections of musical works, represented in symbolic notation, and aims at providing computer-based services on these collections. Let us consider these requirements in turn.

A. Collection of symbolic musical works

A collection is a set of homogeneous musical pieces, e.g., pieces whose contents share some stylistic, structural or notation features. The system should be flexible enough to deal with the inherent subjectivity of this concept, and allow for both intra -and inter- collection analysis and searches. For instance, the two initial collections supported by NEUMA consists of Catholic Liturgic chants (*Sequentia*, see Section IV) and French Renaissance Psalters. In terms of musical content, both consist of monodic chants, and share some historical and structural features whose description is beyond the scope of the present paper. One must therefore envisage exploratory tools that apply at the collection level, or span multiple collections for comparison and analysis purposes.

The notion of symbolic notation covers the content of traditional music scores, put in digital form through a music editing software or via OCR of music sheets. However, in order to be meaningful, this content must be mapped to a semantic-oriented description, a "model" of the music content, apt at supporting a set of relevant but specific functionalities. Defining a general model of music content is probably an overambitious task, and our approach adopts a more realistic (and modest) way.

We consider that a music digital library (MDL) should propose one of several models, each identifying a specific interpretation of a musical piece content in terms of semantically meaningful elements: notes and durations, combination of horizontal (melody) of vertical (chords) combination of notes, synchronization of text and note series, etc. The MDL should

be able to map a digital score representation to an instance of one of the supported models. Well defined functions can then be applied to this instance.

The primary goal of NEUMA is to manage large collections of Western music, and as such we will define models devoted this specific area. The system design should make it possible to address a broader context by incorporating others music abstractions. Managing a collection of folk songs tablatures for instance would probably call for a specific model.

B. Computer-based functionalities for musical content

Given a model that structures the content of a musical work in terms of well-defined elements, the MDL must support content-based functionalities. We decided to focus on the following ones.

Search by content. An application relying on the MDL should be able to express complex content-based searches on music collections. Our intent is to develop each abstraction of music pieces as a database model fully equipped with a query language. In technical terms, this means (among other features) a careful distinction between the physical and the logical level, a set of closed-form operators that can be combined to obtain expressive queries, and a user-oriented, syntax-friendly query language. We are aware that music constitutes an important challenge for such a traditional database approach, because of the unlimited flexibility of its forms and structure. However, by adopting a representation based on symbolic notation, we believe that we can find enough structure in the content to support a semantically significant search tool. The section devoted to the case study illustrates a preliminary attempt in this direction.

Cooperative work based on ontology and annotations. The MDL must allow users to share documents, and also to exchange individual knowledge concerning documents, through an ontology-based annotation system. An ontology is an explicit specification of a conceptualization. Ontological categories define, at a semantic level, the concepts that exist in the domain and relationships between these concepts (in our case music concepts, Catholic Liturgic events, etc.).

Ontology definition is a collective work bringing together ontology experts (called ontologists) and users (musicologists here). For the purpose of NEUMA a referential and global ontology is defined.

An ontology-based tagging system will be included in NEUMA. Annotation systems have been widely studied in the digital libraries community. See for example the Digital Library Annotation Service (DILAS) [1], the Collaboratory for Annotation Indexing and Retrieval of Digitized Historical Archive Material (COLLATE) [2], and the Flexible Annotation Service Tool (FAST) [3]. An annotation consists of a *tag* placed by a user on a document or a part of a document (for a better exploitability, a tag has to belong to the ontology). This enriches the MDL with personal knowledge, and shares this knowledge with other users.

An ontology can also be used to provide a user-friendly GUI[4]. And finally, the ontology can also be exploited during the query process through inferences based on concepts relationships increasing the expressive power of the query language.

Music analysis. This generic term refers to any task that explores the symbolic musical content in order to extract some implicitly represented information. We distinguish these functionalities from query operations because they do not operate in closed-form, rely on statistical evaluations, and involve algorithms which may be specific to a certain subclass of the collections hosted by NEUMA. An interesting feature of computer-driven music analysis is its interdisciplinary nature which implies a close cooperation between musicologists and computer experts. A simple example, currently under investigation on our first collection, is a study of text/music relationships.

A broader ambition is to enable, in the realm of musical content, the analysis and classification tools already in use for text documents, where the content can be automatically analyzed, classified, and indexed by a summary of weighted keywords [5], [6]. Transposed to music content, this means identification of stylistic features, automatic extraction of rythmic or melodic patterns, genre classification, etc.

Copyright protection. A last topic of interest to our project is the protection of copyright. Since the envisioned platform will be able to store and deliver music pieces, or fragments of, we need to address the issue of preserving the ownership of each document. The currently adopted approach is to embed the owner's signature in the musical content without altering the quality of the symbolic representation ("*watermarking*"). This functionality may somehow be considered as independent from those mentioned above, as it could be useful, for instance, from web-based professional sites selling electronic scores.

III. THE ARCHITECTURE OF NEUMA

During the early stages of the project we designed an architecture for NEUMA based on the principles outlined above and capable of fulfilling our requirements. We now describe this architecture and present a first application in the next section.

We adopted the Service Oriented Architecture (SOA) approach because of its ability to combine loosely coupled tools which can be designed and built independently as soon as they agree on a common interface. A global view of the architecture is given in Figure 1. We distinguish two families of components. The first one refers to *Applications* which can be any piece of software, located anywhere on the Internet, that manipulates musical content. The second one constitutes the NEUMA platform which consists of a set of services, each dedicated to a specific functionality. Let us briefly consider them in turn.

The *storage service* allows external applications to register musical documents in NEUMA. A specific abstraction model φ must be chosen by the application, among those proposed by NEUMA (currently only one such model has been implemented: see the next section). The *register()* service takes as input a document d, its global id (used to unambiguously

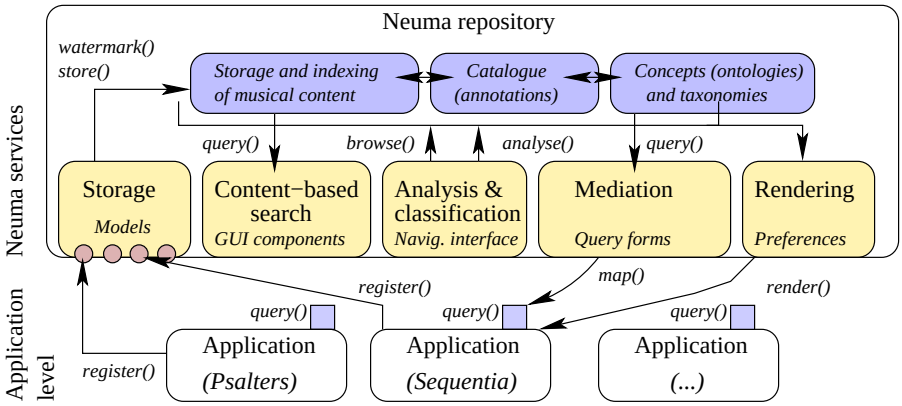


Fig. 1. Architecture of NEUMA

refer to d later on) and extracts from d an instance $i = \varphi(d)$ of the model which is stored in the repository. The set of instances $\{i_1, i_2, \cdots, i_n\}$ collected from an application constitues a *collection* which supports the other services. During this acquisition step, the owner signature can be embedded in instances for copyright protection purposes.

Content-based search is a service that takes a collection of instances and applies transforms (e.g., shifts) or predicates (e.g., musical pattern matching). As mentioned previously, our vision is to develop full-fledged musical query languages that manipulate instances of the model in closed-form and can be composed in arbitrarily complex expressions. This would allow, among other advantages, the definition of a sound query interface for this content-based service, allowing its client applications to submit query expressions to retrieve the matching instances. Our preliminary attempt in this direction is discussed in the next section. The service also provides graphical tools which can be integrated in external applications to facilitate query expression (virtual piano, virtual drums, query by humming, etc.).

The *analysis and classification* service aims at completing the symbolic content representation with annotations. This service is strongly related to the cooperative aspects that we ambition to promote in the NEUMA platform. We will integrate in the digital library tools allowing users to browse, analyse and, ultimately, tag the library content with elements from common vocabularies (or *taxonomies*). The annotation process may be manual, semi-automatic or automatic. A manual approach requires minimal support: the user must be able to browse the music collection, identify elements of interest and tag these elements with terms chosen from an appropriate taxonomy. Semi-automatic and automatic identification requires more involved tools which propose (or even decide) annotations from the result of some content-based analysis.

The *mediation* service supports information retrieval based on *global concepts*. NEUMA will host a set of music-oriented ontologies, general enough to span several collections managed by the digital library, and composed of commonly accepted concepts that can be used to describe musical pieces. We do not envisage to store in the NEUMA repository the instantiation of these concepts for each document. Rather, we will maintain *mappings* between these concepts and the local, application-level, representation. Each application requiring this service needs to register a mapping between its local schema and the NEUMA ontology. This mapping is then used to support description-based search by sending mediated queries (the *query()* local service in Fig. 1) to remote applications. This part of the architecture complies to standard mediation principles (See Section V for a discussion on related works).

The core components of ontological tools should consist of music-centric description: tonality, harmonic sequences, performance indications, etc. We also envisage the development of contextual ontologies that help to interpret and analyse musical pieces with respect to a historical or social environment. These ontologies should be of general value, i.e., they should apply to several collections and support inter-collection search and comparisons.

Finally, the *rendering service* is used to produce a user-friendly representation of the NEUMA services output. Typically, the set of pieces that constitute the result of a search operation must be shown as musical scores (or fragments of). Such results may be created dynamically by, for instance, transposing/merging/slicing data found in the repository.

IV. A CASE STUDY: SEQUENTIA

We implemented a first version of the envisioned architecture to test its ability to fulfil our goals. Our initial targets consist of two collections: Catholic Liturgic chants from the XVIIe century, and Psalms from the same period. Those two collections, SEQUENTIA and PSALTERS, although clearly distinct, share many features that make their association interesting. The description that follows focuses on SEQUENTIA but applies to both collections.

SEQUENTIA can be accessed on-line at *http://sequentia.irpmf-cnrs.fr*. In terms of functionalities, it compares to other similar sites (see for instance *http://www.globalchant.org/*) with a somehow greater search expressivity due to the ability to combine taxonomy-based criteria with similarity content-based retrieval. The

description that follows is meant as an illustration of the architectural principles outlined previously.

The data model. The data model supports the representation of polyphonics pieces (called "documents" hereafter) composed of "voices", each voice being a sequence of notes such that only one note can be played at a given instant. In other words, the model ignores polyphonic instruments. We extend this intuitive concept to any sequence of symbols taken from a finite alphabet. This allows to cover melodies (where the alphabet consists of notes) and text (where the alphabet consists of syllabes) as well as, potentially, any sequence of music-related information (e.g., fingerings).

A *musical piece* p is modeled as a set of synchronized times series $(s_1, s_2, \cdots, s_n)$ that all cover a same time frame $\mathcal{T} = [t_1, t_2]$. The time domain is discrete (the unit being the minimal note duration). Each time serie s_i is identified by a unique label λ_i which can be seen, at an abstract level, as a function $\lambda_i : \mathcal{T} \rightarrow \Sigma$ where Σ is a set of symbols (notes, text, fingerings, etc.) that constitute the "domain" of s_i.

A musical piece can then be viewed as a heterogeneous matrix where lines are instruments (or text) and columns are the sequence of instants. We insert a monophonic melody into a matrix line by converting each note to the numerical value of its pitch. In the SEQUENTIA collection, each piece is a 2-lines matrix where one line is the monophonic melody and the other is the corresponding syllable. Chamber music will have more lines, longer pieces with different movements can be split through several matrix.

The query language can then be conceptualized in terms of matrix operations. Here are a few examples:

- Get one or more instruments from a piece: get lines $\lambda_i, \cdots, \lambda_j$.
- Transpose a piece: add a constant matrix.
- Get all events at instant q: projet on column t_q.
- Get all events from a time slice: get columns $t_q, \cdots, t_{q+k}$.
- Add or remove a movement from a piece: concatenate/subtract two matrixes column-wise.
- Add or remove instruments from a piece: concatenate/subtract two matrixes line-wise.

All these operations operate in closed form and can therefore be combined to form complex expressions. Note that operations on the lines of a matrix can generate information not explicitly featured in the music score. Computing a harmonic progression involves a difference operator between the one that takes several lines as input and the one that produces a new one. Such an algebra constitutes the backbone of a user-friendly query language proposing more meaningful expressions.

Query primitives. The data model is implemented in C++, with serialization functions that allow to put/retrieve a piece to/from a persistent repository. The matrix operations (currently under implementation) are based on a few primitives: subsequence matching, slicing, shifting, etc. The main primitive implemented so far is an approximate pattern matching based on the *Dynamic Time Warping* (DTW) algorithm. It is used for content-based search, either for ranking query results, or during data retrieval (in that case a pre-defined threshold is applied) when a music pattern is provided by the user using a virtual piano (see Fig 2).

Music analysis. We did not investigate automatic music analysis algorithms so far. An ongoing work on our two collections addresses the issue of text-music relationship. The goal is to extract musical fragments associated to text sentences, and to study how the melodic shape matches the meaning of the text.

Taxonomies. The platform (in its current state of implementation) supports a simple mechanism for content annotations based on controlled vocabularies, or *taxonomies*. A taxonomy consists of a set of terms together with a subsumption relation between terms. If $s \preceq t$ then we say that s is *subsumed* by t, or that t *subsumes* s. A taxonomy is usually represented as a graph, where the nodes are the terms and there is an arrow from term s to term t if s subsumes t. Figure 3 shows an excerpt of one of the taxonomies used in SEQUENTIA (description of Liturgic offices). Two terms s and t are *synonyms* if $s \preceq t$ and $t \preceq s$. Given a taxonomy $(T, \preceq)$, an *annotation* in T is any set of terms from T. Users can attach annotations to the documents, and annotations can be used as a support to find and retrieve the document.

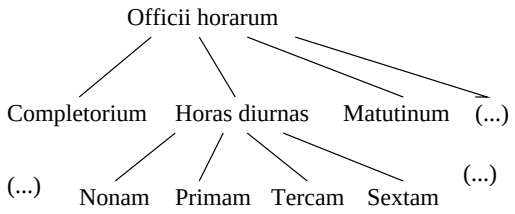


Fig. 3. A taxonomy (excerpt)

The annotation of a d is stored in a repository which can be thought of as a set of pairs (t, d) for each term t appearing in the annotation of d. The set of all such pairs (t, d), for all documents is the *catalogue* (see Fig. 1).

A *query* over the catalogue is any string derived by the following grammar, where t is a term:

$$q ::= t | q \wedge q' | q \vee q' | q \wedge \neg q' | (q)$$

If the query is a single term, then the answer is the set of all documents related either to t or to a term subsumed by t. If the query is not a single term then we proceed as follows. First, for each term appearing in the query, replace the term by the set of all documents computed as explained above; then replace each boolean combinator appearing in the query by the corresponding set-theoretic operator; finally, perform the set-theoretic operations to find the answer.

SEQUENTIA uses on three taxonomies: *Offices*, *Liturgic calendar* and *Solemnity level*. Term-based queries can be submitted via a web form that allows to select the terms of interest to retrieve a set of documents.

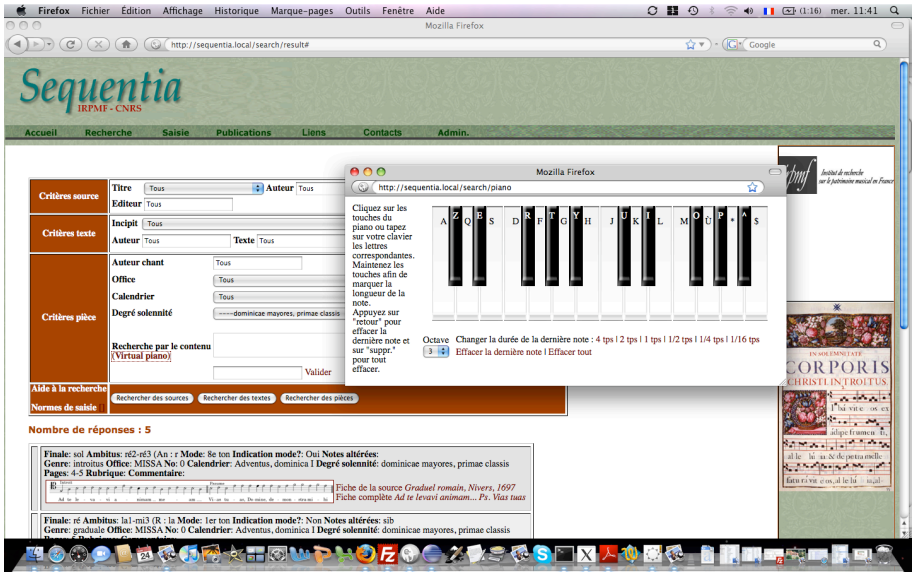


Fig. 2. The search form of SEQUENTIA

Current state of implementation and discussion.

Figure 2 shows the main query form in SEQUENTIA, along with a few results. The form allows to combine content-based search (melodic fragments can be entered with a virtual piano) and term-based search. The list of music pieces matching the query criteria is displayed as a ranked list, each item being a link to navigate in the collection.

The *register()* function consists in sending a MusicXML containing the description of a new musical document. A mapping function extracts from the XML representation an instance of our data model which is then stored in the repository. This document can be annotated with terms from the Catalogue. MusicXML (*http://www.recordare.com*) is now a widely used exchange format for musical symbolic content, supported by almost all the music editor softwares. Other symbolic formats could be envisaged as well, providing that the appropriate extraction function is supplied.

The *render()* function relies on Lilypond (*http://lilypong.org*). Currently the score is produced during the registration of a new piece, and cannot be changed later on. This prevent the dynamic generation of scores from query results. We plan to devote a part of the project resources to implement a convenient API upon Lilypond to obtain a more flexible and powerful tool.

As mentioned above, our primary goal was to validate the main lines of our approach. The services that compose our envisioned architecture have been implemented with basic functionalities, and interact to support this simple application. In terms of query language and query evaluation techniques, the current implementation lacks from efficient mechanisms such as, for instance, indexing of musical fragments. However, the combination of approximate content-based retrieval on one side, with the term-based query language built on taxonomies on the other side, turns out to provide a useful search tool and constitutes an encouraging result for further development of the NEUMA components. In particular, we are now working on a stronger querying service based on a query algebra, powerful primitive implementations, and a user query language that isolates the application from the internal mechanisms. Access paths to efficiently support common searches on large music collections is also underway [7], [8].

V. RELATED WORK

The past decade has witnessed a growing interest in techniques for representing, indexing and searching (by content) music documents. The domain is commonly termed "Music Information Retrieval" (MIR) although it covers many aspects beyond the mere process of retrieving documents. We refer the reader to [9] for an introduction. Roughly speaking, systems can manipulate music either as audio files or in symbolic form. NEUMA belongs to the latter category. The symbolic representation offers a structured representation which is well suited for content-based accesses and sophisticated manipulations and analysis [10].

Several projects have been or are currently devoted to MIR and digital libraries. Close projects to NEUMA are the OMRAS (Online Music Recognition and Searching)[11] and OMRAS2 (Ontology-driven Music Retrieval and Annotation Sharing Service) [12] projects. OMRAS's ambition was to build a system for content-based searching of online musical databases via an intuitive interface that uses music in a visual or aural form. OMRAS paid a particular attention to music transcription that might be seen on a score [13], [14]. OMRAS2 is a framework for annotating and searching collections of both recorded music and digital score representations (like MIDI). Of particular interest to our goals are the logic-based representation and management [15], [16] and social networks analysis [17], [18].

Architecture. The architecture of NEUMA follows the current trend of building structured views over loosely structured doc-

uments. A representative example is the COUCHDB Apache system [19] which supports the definition and manipulation of views built using the Map/Reduce paradigm. A NEUMA model can be seen as a view implementing a particular interpretation of some musical content.

The mediation approach has been widely studied in databases and AI domains (see [20] for an overview). A mediator provides a uniform query interface for querying collections of pre-existing data sources that were created independently. Two approaches of mediation can be considered, depending on the way mappings are defined: the GAV (Global As Views) and the LAV (Local As Views) approach. A comparison of this approaches can be found in [21]. Number of mediator systems exist (e.g. PICSEL [22], Xyleme [23], TSIMMIS[24], etc).

Ontologies. Ontologies are means to describe agreed and shared knowledge for communities of people. Different levels of knowledge representation can be considered in the specification of an ontology, ranging from *taxonomies* to define an agreed set of terms used in a specific domain, to more complex conceptual models that can interact as mediation models in a cooperation of information retrieval systems. The Music Ontology [25] addresses a large panel of knowledge including editorial, cultural and acoustic information. The MPEG-7 ontologies [26] deal to model standard meta-data to interlink different multimedia resources. The MX ontology [27] is more related to music description and classification. Ontologies in NEUMA ambition to provide a support model to interlink several musical collections.

VI. CONCLUSION

NEUMA aims at exploring the range of possibilities offered by the application of well established digital libraries services to large archives of symbolic musical content. The project roadmap is based on the step-by-step development of functionalities supporting several collections, ranging from simple monodic pieces to large and complex musical works, e.g., piano concerti.

We are aware that music is inherently a quite various and unpredictible material, and the design of our platform must be flexible enough to cope with it. Our ambition is to produce results of interest to the community of users as well as to contribute to advances in the realm of mutimedia digital libraries.

ACKNOWLEDGMENTS

This work is supported by the French ANR project NEUMA (*http://neuma.irpmf-cnrs.fr*).

REFERENCES

[1] M. Agosti, H. Albrechtsen, N. Ferro, I. Frommholz, P. Hansen, N. Orio, E. Panizzi, A. M. Pejtersen, and U. Thiel, "Dilas: a digital library annotation service," in *IWAC*, J.-F. Boujut, Ed. CNRS - Programme société de l'information, 2005, pp. 91–101.

[2] J. Keiper, H. Brocks, A. Dirsch-Weigand, A. Stein, and U. Thiel, "Collate - a web-based collaboratory for content-based access to and work with digitized cultural material," in *ICHIM (1)*, 2001, pp. 495–511.

[3] M. Agosti and N. Ferro, "A system architecture as a support to a flexible annotation service," in *DELOS Workshop: Digital Library Architectures - LNCS Volume*, ser. Lecture Notes in Computer Science, C. Türker, M. Agosti, and H.-J. Schek, Eds., vol. 3664. Springer, 2004, pp. 147–166.

[4] L. Cinque, A. Malizia, and R. Navigli, "Ontodoc: An ontology-based query system for digital libraries," in *ICPR (2)*, 2004, pp. 671–674.

[5] K. S. Jones and D. M. Jackson, "The use of automatically-obtained keyword classifications for information retrieval," *Information Storage and Retrieval*, vol. 5, no. 4, pp. 175–201, 1970.

[6] M. J. Dovey, "Adding content-based searching to a traditional music library catalogue server," in *Proc. Joint Conf. on Digital Libraries (JCDL)*, 2001, pp. 249–250.

[7] Y. Zhu and D. Shasha, "Warping Indexes with Envelope Transforms for Query by Humming," in *Proc. ACM SIGMOD Symp. on the Management of Data*, 2003, pp. 181–192.

[8] E. J. Keogh and C. A. Ratanamahatana, "Exact Indexing of Dynamic Time Warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, 2005.

[9] M. Muller, *Information Retrieval for Music and Motion*. Springer, 2004.

[10] G. Haus, M. Longari, and E. Pollstri, "A Score-Driven Approach to Music Information Retrieval," *Journal of American Society for Information Science and Technology*, vol. 55, pp. 1045–1052, 2004.

[11] "Online Music Recognition and Searching (OMRAS) project," http://www.omras.org.

[12] "Ontology-driven Music Retrieval and Annotation Sharing Service (OMRAS2) project," http://www.omras2.org.

[13] J. P. Bello, G. Monti, and M. B. Sandler, "Techniques for automatic music transcription," in *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, 2000.

[14] J. Bello, L. Daudet, and M. Sandler, "Time-domain polyphonic transcription using self-generating databases," in *Proc. of the 112th Convention of the Audio Engineering Society*, 2002.

[15] A. Anglade and S. Dixon, "Characterisation of harmony with inductive logic programming," in *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, 2008.

[16] ——, "Towards logic-based representations of musical harmony for classification, retrieval and knowledge discovery," in *MML*, 2008.

[17] A. Anglade, M. Tiemann, and F. Vignoli, "Virtual communities for creating shared music channels," in *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, 2007.

[18] K. Jacobson, B. Fields, and M. Sandler, "Using audio analysis and network structure to identify communities in on-line social networks of artists," in *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, 2008.

[19] J. C. Anderson, N. Slater, and J. Lehnardt, *CouchDB: Rough Cuts Version*. O'Reilly Editions, 2008.

[20] G. Wiederhold, "Mediators in the architecture of future information systems," *IEEE Computer*, vol. 25, no. 3, pp. 38–49, 1992.

[21] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching," *VLDB J.*, vol. 10, no. 4, pp. 334–350, 2001.

[22] F. Goasdoué, V. Lattès, and M.-C. Rousset, "The use of carin language and algorithms for information integration: The picsel system," *Int. J. Cooperative Inf. Syst.*, vol. 9, no. 4, pp. 383–401, 2000.

[23] A. Y. Levy, A. Rajaraman, and J. J. Ordille, "Query-answering algorithms for information agents," in *AAAI/IAAI, Vol. 1*, 1996, pp. 40–47.

[24] S. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. D. Ullman, and J. Widom, "The tsimmis project: Integration of heterogeneous information sources," in *IPSJ*, 1994, pp. 7–18.

[25] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson, ""the music ontology"," in *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, 2007, pp. 417–422.

[26] R. Troncy, S. Celma, O.and Little, R. Gracia, and C. Tsinaraki, "Mpeg-7 based multimedia ontologies: Interoperabity support or interoperability issue?" in *Proc. of the 1st Worshop on Multimedia Annotation and Retrieval enabled by shared Ontologies (MAReSO'07) in conjunction with SAMT'07*, 2007, pp. 2–14.

[27] A. Ferrara, L. A. Ludovico, S. Montanelli, S. Castano, and G. Haus, "A semantic web ontology for context-based classification and retrieval of music resources," *TOMCCAP*, vol. 2, no. 3, pp. 177–198, 2006.

Analytic Comparison of Audio Feature Sets using Self-Organising Maps

Rudolf Mayer, Jakob Frank, Andreas Rauber
Institute of Software Technology and Interactive Systems
Vienna University of Technology, Austria
{mayer,frank,rauber}@ifs.tuwien.ac.at

***Abstract*—A wealth of different feature sets for analysing music has been proposed and employed in several different Music Information Retrieval applications. In many cases, the feature sets are compared with each other based on benchmarks in supervised machine learning, such as automatic genre classification. While this approach makes features comparable for specific tasks, it doesn't reveal much detail on the specific musical characteristics captured by the single feature sets. In this paper, we thus perform an analytic comparison of several different audio feature sets by means of Self-Organising Maps. They perform a projection from a high dimensional input space (the audio features) to a lower dimensional output space, often a two-dimensional map, while preserving the topological order of the input space. Comparing the stability of this projection allows to draw conclusions on the specific properties of the single feature sets.**

I. Introduction

One major precondition for many Music Information Retrieval (MIR) tasks is to adequately describe music, resp. its sound signal, by a set of (numerically processable) feature vectors. Thus, a range of different audio features has been developed, such as the Mel-frequency cepstral coefficients (MFCC), the set of features provided by the MARSYAS system, or the Rhythm Patterns, Rhythm Histograms and Statistical Spectrum Descriptors suite of features.

All these feature sets capture certain different characteristics of music, and thus might perform unequally well in different MIR tasks. Very often, feature sets are compared by the means of benchmarks, e.g. the automated classification of music towards a certain label, such as in automatic genre classification. While this allows a comparative evaluation of different feature sets with respect to specific tasks, it doesn't provide many insights on the properties of each feature set. On the other hand, clustering or projection methods can reveal information such as which data items tend to be organised together, revealing information on the acoustic similarities captured by the respective feature sets. Building on this assumption, we utilise a recently developed method to compare different instances of a specific projection and vector quantisation method, the Self-Organising Maps, to compare how the resulting map is influenced by the different feature sets.

The remainder of this paper is structured as follows. Section II discusses related work in Music Information Retrieval and Self-Organising Maps, while Section III presents the employed audio features in detail. Section IV then introduces the method for comparing Self-Organising Maps. In Section V we introduce the dataset used, and discuss experimental results. Finally, Section VI gives conclusions and presents future work.

II. Related Work

Music Information Retrieval (MIR) is a discipline of Information Retrieval focussing on adequately describing and accessing (digital) audio. Important research directions include, but are not limited to, similarity retrieval, musical (genre) classification, or music analysis and knowledge representation.

The dominant method of processing audio files in MIR is by analysing the audio signal. A wealth of different descriptive features for the abstract representation of audio content have been presented. The feature sets we used in our experiments, i.e. *Rhythm Patterns* and derived sets, *MARSYAS*, and *Chroma*, are well known algorithms focusing on different audio characteristics and will be described briefly in Section III.

The Self-Organising Map (SOM) [1] is an artificial neural network used for data analysis in numerous applications. The SOM combines principles of vector projection (mapping) and vector quantisation (clustering), and thus provides a mapping from a high-dimensional input space to a lower dimensional output space. The output space consists of a certain number of nodes (sometimes also called units or models), which are often arranged as a two-dimensional grid, in rectangular or hexagonal shape. One important property of the SOM is the fact that it preserves the topology of the input space as faithfully as possible, i.e. data that is similar and thus close to each other in the input space will also be located in vicinity in the output map. The SOM thus can be used to uncover complex inherent structures and correlations in the data, which makes it an attractive tool for data analysis.

The SOM has been applied in many Digital Library settings, to provide a novel, alternative way for browsing the library's content. This concept has also been applied for Music Retrieval to generate *music maps*, such as in the SOMeJB [2] system. Specific domain applications of music maps are for example the *Map of Mozart* [3], which organises the complete works of Mozart in an appealing manner, or the *Radio SOM* [4], illustrating musical profiles of radio stations. A comprehensive overview on music maps, with a special focus on the user interaction with them, can be found in [5].

III. Audio Features

In our experiments, we employ several different sets of features extracted from the audio content of the music, and compare them to each other. Specifically, we use the MARSYAS, Chroma, Rhythm Patterns, Statistical Spectrum Descriptors, and Rhythm Histograms audio feature sets, all of which will be described below.

A. MARSYAS Features

The MARSYAS system [6] is a software framework for audio analysis, feature extraction and retrieval. It provides a number of feature extractors that can be divided into three groups: features describing the timbral texture, those capturing the rhythmic content, and features related to pitch content.

The *STFT-Spectrum based Features* provide standard temporal and spectral low-level features, such as Spectral Centroid, Spectral Rolloff, Spectral Flux, Root Mean Square (RMS) energy and Zero Crossings. Further, MARSYAS computes the first twelve Mel-frequency cepstral coefficients (MFCCs).

The rhythm-related features aim at representing the regularity of the rhythm and the relative saliences and periods of diverse levels of the metrical hierarchy. They are based on the *Beat Histogram*, a particular rhythm periodicity function representing beat strength and rhythmic content of a piece of music. Various statistics are computed of the histogram: the relative amplitude of the first and second peak, the ratio of the amplitude of the second peak and the first peak, the period of the first and second beat (in beats per minute), and the overall sum of the histogram, as indication of beat strength.

The *Pitch Histogram* is computed by decomposing the signal into two frequency bands, for each of which amplitude envelopes are extracted and summed up, and the main pitches are detected. The three dominant peaks are accumulated into the histogram, containing information about the pitch range of a piece of music. A folded version of the histogram, obtained by mapping the notes of all octaves onto a single octave, contains information about the pitch classes or the harmonic content. The amplitude of the maximum peak of the folded histogram (i.e. magnitude of the most dominant pitch class), the period of the maximum peak of the unfolded (i.e. octave range of the dominant pitch) and folded histogram (i.e. main pitch class), the pitch interval between the two most prominent peaks of the folded histogram (i.e. main tonal interval relation) and the overall sum of the histogram are computed as features.

B. Chroma Features

Chroma features [7] aim to represent the harmonic content (e.g, keys, chords) of a short-time window of audio by computing the spectral energy present at frequencies that correspond to each of the 12 notes in a standard chromatic scale (e.g., black and white keys within one octave on a piano). We employ the feature extractor implemented in the MARSYAS system, and compute four statistical values for each of the 12-dimensional Chroma features, thus resulting in a 48-dimensional feature vector.

C. Rhythm Patterns

Rhythm Patterns (RP) are a feature set for handling audio data based on analysis of the spectral audio data and psycho-acoustic transformations [8], [9].

In a pre-processing stage, multiple channels are averaged to one, and the audio is split into segments of six seconds, possibly leaving out lead-in and fade-out segments.

The feature extraction process for a Rhythm Pattern is then composed of two stages. For each segment, the spectrogram of the audio is computed using the short time Fast Fourier Transform (STFT). The window size is set to 23 ms (1024 samples) and a Hanning window is applied using 50 % overlap between the windows. The Bark scale, a perceptual scale which groups frequencies to critical bands according to perceptive pitch regions [10], is applied to the spectrogram, aggregating it to 24 frequency bands. Then, the Bark scale spectrogram is transformed into the decibel scale, and further psycho-acoustic transformations are applied: computation of the Phon scale incorporates equal loudness curves, which account for the different perception of loudness at different frequencies [10]. Subsequently, the values are transformed into the unit Sone. The Sone scale relates to the Phon scale in the way that a doubling on the Sone scale sounds to the human ear like a doubling of the loudness. This results in a psycho-acoustically modified Sonogram representation that reflects human loudness sensation.

In the second step, a discrete Fourier transform is applied to this Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band. After additional weighting and smoothing steps, a Rhythm Pattern exhibits magnitude of modulation for 60 modulation frequencies (between 0.17 and 10 Hz) on 24 bands, and has thus 1440 dimensions.

In order to summarise the characteristics of an entire piece of music, the feature vectors derived from its segments are averaged by computing the median.

D. Statistical Spectrum Descriptors

Computing Statistical Spectrum Descriptors (SSD) features relies on the first stage of the algorithm for computing RP features. Statistical Spectrum Descriptors are based on the Bark-scale representation of the frequency spectrum. From this representation of perceived loudness, a number of statistical measures is computed per critical band, in order to describe fluctuations within the critical bands. Mean, median, variance, skewness, kurtosis, min- and max-value are computed for each of the 24 bands, and a Statistical Spectrum Descriptor is extracted for each selected segment. The SSD feature vector for a piece of audio is then calculated as the median of the descriptors of its segments.

In contrast to the Rhythm Patterns feature set, the dimensionality of the feature space is much lower – SSDs have $24 \times 7 = 168$ instead of 1440 dimensions – at matching performance in terms of genre classification accuracies [9].

E. Rhythm Histogram Features

The Rhythm Histogram features are a descriptor for the rhythmic characteristics in a piece of audio. Contrary to the Rhythm Patterns and the Statistical Spectrum Descriptor, information is not stored per critical band. Rather, the magnitudes of each modulation frequency bin (at the end of the second phase of the RP calculation process) of all 24 critical bands are summed up, to form a histogram of 'rhythmic energy' per modulation frequency. The histogram contains 60 bins which reflect modulation frequency between 0.168 and 10 Hz. For a given piece of audio, the Rhythm Histogram feature set is calculated by taking the median of the histograms of every 6 second segment processed.

IV. Comparison of Self-Organising Maps

Self-Organising Maps can differ from each other depending on a range of various factors: simple ones such as different initialisations of the random number generator, to more SOM-specific ones such as different parameters for e.g. the learning rate and neighbourhood kernel (cf. [1] for details), to differences in the map-size. In all such cases, the general topological ordering of the map should stay approximately the same, i.e. clusters of data items would stay in the neighbourhood of similar clusters, and be further away from dissimilar ones, unless the parameters were chosen really bad. Still, some differences will appear, which might then range from e.g. a minor deviation such as a mirrored arrangement of the vectors on the map, to having still the same local neighbourhood between specific clusters, but a slightly rotated or skewed global layout. Training several maps with different parameters and then analysing the differences can thus give vital clues on the structures inherent in the data, by discovering which portions of the input data are clustered together in a rather stable fashion, and for which parts random elements play a vital role for the mapping.

An analytic method to compare different Self-Organising Maps, created with such different training parameters, but also with different sizes of the output space, or even with different feature sets, has been proposed in [11]. For the study presented in this paper, especially the latter, comparing different feature sets, is of major interest. The method allows to compare a selected *source* map to one or more *target* maps by comparing how the input data items are arranged on the maps. To this end, it is determined whether data items located close to each other in the source map are also closely located to each other in the target map(s), to determine whether there are *stable* or *outlier* movements between the maps. "Close" is a user-adjustable parameter, and can be defined to be on the same node, or within a certain radius around the node. Using different radii for different maps accommodates for maps differing in size. Further, a higher radius allows to see a more abstract, coarse view on the data movement. If the majority of the data items stays within the defined radius, then this is regarded a stable shift, or an outlier shift otherwise. Again, the user can specify how big the percentage needs to be to regard it a stable or outlier shift. These shifts are visualised by arrows, where different colours indicate stable or outlier shifts, and the line-width determines the cardinality of the data items moving along the shift. The visualisation is thus termed *Data Shifts* visualisations. Figure 1 illustrates stable (green arrows) and outlier (red arrows) shifts on selected nodes of two maps, the left one trained with Rhythm Pattern, the right one with SSD features. Already from this illustration, we can see that some data items will also be closely located on the SSD map, while others spread out to different areas of the map.

Finally, all these analysis steps can be done as well not on a per-node basis, but rather regarding clusters of nodes instead. To this end, first a clustering algorithm is applied to the two maps to be compared to each other, to compute the same, user-adjustable number of clusters. Specifically, we use Ward's linkage clustering [12], which provides a hierarchy of clusters at different levels. The best-matching clusters found in both SOMs are then linked to each other, determined by the highest matching number of data points for pairs of clusters on both maps – the more data vectors from cluster A_i in the first SOM are mapped into cluster B_j in the second SOM, the higher the confidence that the two clusters correspond to each other. Then all pairwise confidence values between all clusters in the maps are computed. Finally, all pairs are sorted and repeatedly the match with the highest values is selected, until all clusters have been assigned exactly once. When the matching is determined, the *Cluster Shifts* visualisation can easily be created, analogous to the visualisation of Data Shifts.

An even more aggregate and abstract view on the input data movement can be provided by the *Comparison Visualisation*, which further allows to compare one SOM to several other maps in the same illustration. To this end, the visualisation colours each unit u in the main SOM according to the average pairwise distance between the unit's mapped data vectors in the other s SOMs. The visualisation is generated by first finding all k possible pairs of the data vectors on u, and compute the distances d_{ij} of the pair's positions in the other SOMs. These distances are then summed and averaged over the number of pairs and the number of compared SOMs, respectively. Alternatively to the mean, the variance of the distances can be used.

V. Analytic Comparison of Audio Feature Sets

In this section, we outline the results of our study on comparing the different audio feature sets with Self-Organising Maps.

A. Test Collection

We extracted features for the collection used in the ISMIR 2004 genre contest[1], which we further refer to as *ISMIRgenre*. The dataset has been used as benchmark for several different MIR systems. It comprises 1458 tracks, organised into six different genres. The greatest part of the tracks belongs to Classical music (640, colour-coded in red), followed by World (244, cyan), Rock/Pop (203, magenta), Electronic (229, blue), Metal Punk (90, yellow), and finally Jazz/Blues (52, green).

[1] http://ismir2004.ismir.net/ISMIR_Contest.html

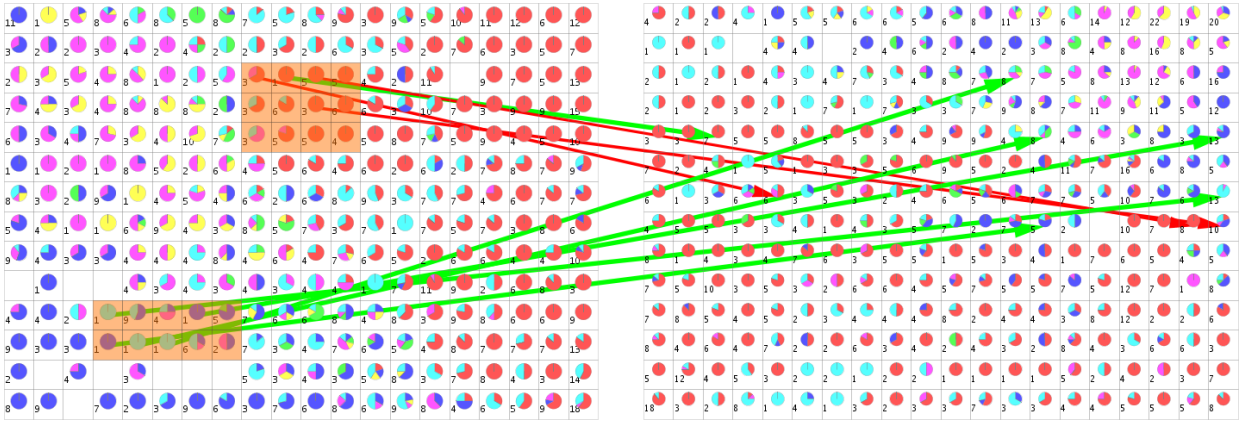


Fig. 1. Data Shifts Visualisation for RP and SSD maps on the ISMIRgenre data sets

TABLE I
CLASSIFICATION ACCURACIES ON THE ISMIRGENRE DATABASES.

Feature Set	1-nn	3-nn	Naïve B.	SVM
Chroma	39.59	45.54	40.73	*45.07*
Rhythm Histograms	60.45	63.04	56.74	*63.74*
MARSYAS	66.69	64.36	59.00	*67.02*
Rhythm Patterns	73.21	71.37	63.31	*75.17*
SSD	**78.20**	**76.44**	**60.88**	***78.73***

B. Genre Classification Results

To give a brief overview on the discriminative power of the audio feature sets, we performed a genre classification on the collection, using the WEKA machine learning toolkit[2]. We utilised k-Nearest-Neighbour, Naïve Bayes and Support Vector Machines, and performed the experiments based on a ten-fold cross-validation, which is further averaged over ten repeated runs. The results given in Table I are the micro-averaged classification accuracies.

There is a coherent trend across all classifiers. It can be noted that SSD features are performing best on each single classifier (indicated by bold print), achieving the highest value with Support Vector Machines, followed surprisingly quite closely by 1-nearest-neighbours. Also the subsequent ranks don't differ across the various classifiers, with Rhythm Patterns being the second-best feature sets, followed by MARSYAS, Rhythm Histograms and the Chroma features. In all cases, SVM are the dominant classifier (indicated by italic type), with the k-NN performing not that far off of them. These results are in line with those previously published in the literature.

C. Genre Clustering with Music Maps

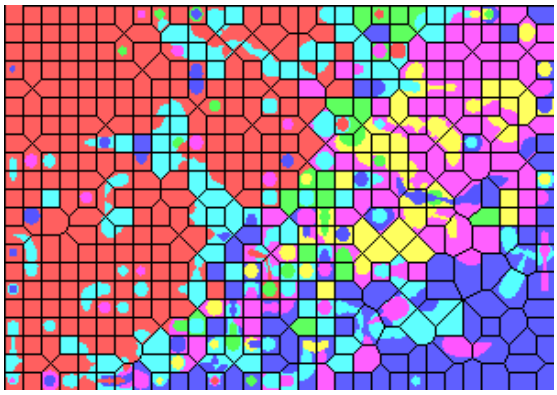
We trained a number of Self-Organising Maps, with different parameters for the random number generator, the number of training iterations, and in different size, for each of the five feature sets. An interesting observation is the arrangement of the different genres across the maps, which is illustrated in Figure 2. While the different genres form pretty clear and distinct clusters on RP, RH and SSD features this is not so much the case for Chroma or MARSYAS features. Figure 2(a) shows the map on RP features. It can be quickly observed that the genres Classical (red), Electronic (blue) and Rock/Pop (magenta) are clearly arranged closely to each other on the map; also Metal/Punk (yellow) and Jazz/Blues (green) are arranged on specific areas of the map. Only World Music (cyan) is spread over many different areas; however, World Music is rather a collective term for many different types of music, thus this behaviour seems not surprising. The maps for RH and SSD features exhibit a very similar arrangement.

For the MARSYAS maps, a pre-processing step of normalising the single attributes was needed, as otherwise, different value ranges of the single features would have a distorting impact on distance measurements, which are an integral part of the SOM training algorithm. We tested both a *standard score* normalisation (i.e. subtracting the mean and dividing by the standard deviation) and a min-max normalisation (i.e. values of range [0..1] for each attribute). Both normalisation methods dramatically improved the subjective quality of the map, both showing similar results. Still, the map trained with the MARSYAS features, depicted in Figure 2(b), shows a less clear clustering according to the pre-defined genres. The Classical genre occupies a much larger area, and is much more intermingled with other genres, and actually divided in two parts by genres such as Rock/Pop and Metal/Punk. Also, the Electronic and Rock/Pop genres are spread much more over the map than with the RP/RH/SSD features. A subjective evaluation by listening to some samples of the map also found the RP map to be superior in grouping similar music.

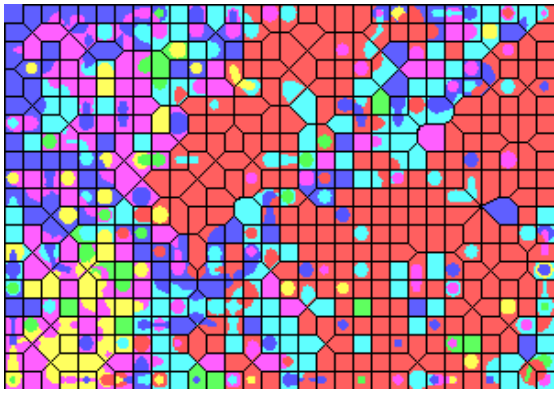
Similar observations hold also true for all variations of parameters and sizes trained, and can further be observed for maps trained on Chroma features.

Thus, a first surprising finding is that MARSYAS features, even though they provide good classification results, outperforming RH features on all tested classifiers and not being that far off from the results with RP, are not exhibiting properties that would allow the SOM algorithm to cluster them as well as with the other feature sets.

[2]http://www.cs.waikato.ac.nz/ml/weka/

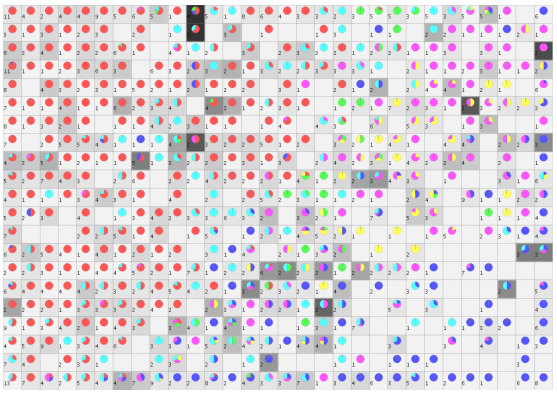


(a) Rhythm Patterns

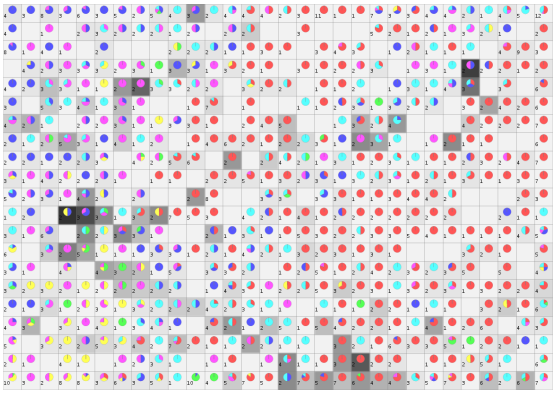


(b) MARSYAS

Fig. 2. Distribution of genres on two maps trained with the same parameters, but different feature sets



(a) Rhythm Patterns



(b) MARSYAS

Fig. 3. Comparison of the two maps from Figure 2 to other maps trained on the same respective feature set, but with different training parameters

D. Mapping Stability

Next, we present the analysis of the stability of the mapping on single feature sets, i.e., we compare maps trained with the same feature sets, but different parameters, to each other. One such visualisation is depicted in Figure 3(a), which compares maps trained with RP features. The darker the nodes on the map, the more instable the mapping of the vectors assigned to these nodes is in regard to the other maps compared to. We can see that quite a big area of the map seems to be pretty stable in mapping behaviour, and there are just a few areas that get frequently shuffled on the map. Most of those are in areas that are the borderlines between clusters that each contain music from a specific genre. Among those, an area in the upper-middle border of the map holds musical pieces from Classical, Jazz/Blues, Electronic, and World Music genres. Two areas, towards the right-upper corner, are at intersections of Metal/Punk and Pop/Rock genres, and frequently get mapped into slightly different areas on the map. We further trained a set of smaller maps, on which we observed similar patterns.

While the SOMs trained with the MARSYAS features are not preserving genres topologically on the map, the mapping itself seems to be stable, as can be seen in Figure 3(b). From a visual inspection, it seems there are not more "instable" areas on the map than with the RP features, and as well, they can be mostly found in areas where genre-clusters intermingle.

E. Feature Comparison

Finally, we want to compare maps trained on different feature sets. Figure 4 shows a comparison of an RP with an SSD map, both of identical size. The Rhythm Patterns map is expected to cover both rhythm and frequency information from the music, while the Statistical Spectrum Descriptors are only containing information on the power spectrum. Thus, an increased number of differences in the mapping is expected when comparing these two maps, in contrast to a comparison of maps trained with the same feature set. This hypothesis is confirmed by a visual inspection of the visualisation, which shows an increased amount of nodes colour-coded to have high mapping distances in the other map.

Those nodes are the starting point for investigating how the pieces of music get arranged on the maps. In Figure 4, a total of four nodes, containing two tracks each, have been selected in the left map, trained with the RP features. In the right map, trained with the SSD features, the grouping of the tracks is different, and no two tracks got matched on the same node or even neighbourhood there. Rather, from both the lower-leftmost and upper-rightmost node containing Classical music, one track each has been grouped together closely at the centre-right area and at the left-centre border. Likewise, the other two selected nodes, one containing World Music, the other World Music and Classical Music, split up in a similar fashion. One

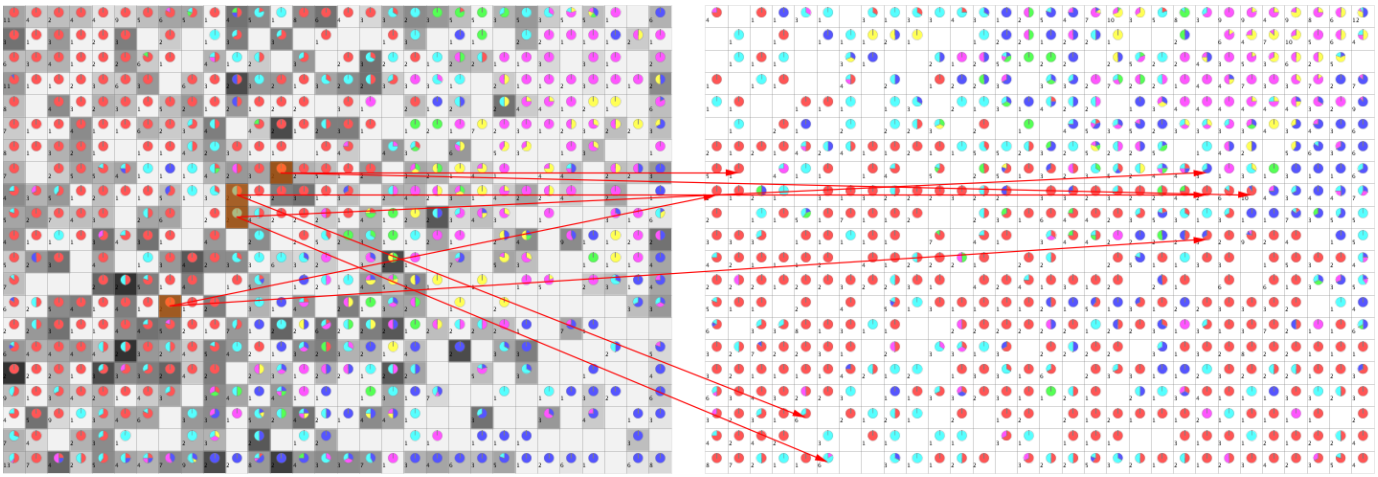


Fig. 4. Comparison of a RP and an SSD map

track each gets mapped to the lower-left corner, at the border of the Classical and World Music cluster. The other two tracks lie in the centre-right area, close two the other two tracks mentioned previously.

Manually inspecting the new clustering of the tracks on the SSD based maps reveals that in all cases, the instrumentation is similar within all music tracks. However, on the RP map, the music is as well arranged by the rhythmic information captured by the feature set. Thus the tracks located on the same node in the left map also share similar rhythmic characteristics, while this is not necessarily the case for the right, SSD-based map.

To illustrate this in more detail, one of the Classical music pieces in the lower-left selected node on the RP map is in that map located clearly separated from another Classical piece on the upper one of the two neighbouring selected nodes in the centre. Both tracks exhibit the same instrumentation, a dominant violin. However, the two songs differ quite strongly in their tempo and beat, the latter music piece being much more lively, while the first has a much slower metre. This differntiates them in the Rhythm Pattern map. However, in the Statistical Spectrum Descriptors, rhythmic characteristics are not considered in the feature set, and thus these two songs are correctly placed in close vicinity of each other.

Similar conclusions can be drawn for comparing other feature sets. An especially interesting comparison is Rhythm Patterns vs. a combination of SSD and Rhythm Histogram features, which together cover very similar characteristics as the Rhythm Patterns, but still differ e.g. in classification results. Also, comparing Rhythm Patterns or Histograms to MARSYAS offers interesting insights, as they partly cover the same information about music, but also have different features.

VI. Conclusions and Future Work

In this paper, we utilised Self-Organising Maps to compare five different audio feature sets regarding their clustering characteristics. One interesting finding was that maps trained with MARSYAS features are not preserving the pre-defined ordering into genres as well as it is the case with RP, RH and SSD features, even though they are similar in classification performance. Further, we illustrated that the approach of using Self-Organising Maps for an analytical comparison of the feature sets can provide vital clues on which characteristics are captured by the various audio feature sets, by highlighting which pieces of music are interesting for a closer inspection. One challenge for future work is to automate detecting interesting patterns as presented in the previous Section.

References

[1] T. Kohonen, *Self-Organizing Maps*, ser. Springer Series in Information Sciences. Berlin, Heidelberg: Springer, 1995, vol. 30.

[2] A. Rauber, E. Pampalk, and D. Merkl, "The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models," *Journal of New Music Research*, vol. 32, no. 2, June 2003.

[3] R. Mayer, T. Lidy, and A. Rauber, "The map of Mozart," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, October 8-12 2006.

[4] T. Lidy and A. Rauber, "Visually profiling radio stations," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Victoria, Canada, October 8-12 2006.

[5] J. Frank, T. Lidy, E. Peiszer, R. Genswaider, and A. Rauber, "Creating ambient music spaces in real and virtual worlds," *Multimedia Tools and Applications*, vol. 44, no. 3, 2009.

[6] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organized Sound*, vol. 4, no. 30, 2000.

[7] M. Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, 2006.

[8] A. Rauber, E. Pampalk, and D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles," in *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'02)*, Paris, France, October 13-17 2002.

[9] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, London, UK, September 11-15 2005.

[10] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, 2nd ed., ser. Series of Information Sciences. Berlin: Springer, 1999, vol. 22.

[11] R. Mayer, D. Baum, R. Neumayer, and A. Rauber, "Analytic comparison of self-organising maps," in *Proceedings of the 7th Workshop on Self-Organizing Maps (WSOM'09)*, St. Augustine, Fl, USA, June 8-10 2009.

[12] J. H. Ward, Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, March 1963.

Matching Places of Interest With Music

Marius Kaminskas
Free University of Bozen-Bolzano
Bolzano, 39100, Italy
Email: mkaminskas@unibz.it

Francesco Ricci
Free University of Bozen-Bolzano
Bolzano, 39100, Italy
Email: fricci@unibz.it

***Abstract*—In this paper we address a particular kind of cross domain personalization task consisting of selecting simultaneously two items in two different domains and recommending them together because they fit the user preferences and also they fit well together. We show that given some personalized recommendations for places of interests (POIs), the user satisfaction for these POIs can be increased by enriching their presentation with music tracks that match the user's profile and are also matching the POIs. We present the results of an online experiment where alternative approaches for matching POIs and music, based on tagging and text matching, have been tested with users.**

I. Introduction

Recommender systems are personalized information search and decision support tools that are generally designed to providing recommendations for just one type of information items, e.g., either movies, or CDs, or travels. Recently, there have been some attempts to combine information from different domains, introducing the notion of cross-domain recommender systems. These systems can reuse knowledge about the users, which is derived in one domain, e.g., CDs, to provide recommendations in another different domain, e.g., movies [1].

In this paper we focus on a particular kind of cross-domain recommendation task, not yet well studied in the literature, where we try to select simultaneously two items in two different domains and to recommend them together because they fit the user preferences and they fit well together. Consider for instance a person who is sightseeing a city. The tourist may use such a recommender system running on her mobile device. This system can offer to the user a walking itinerary and, while the user is visiting the suggested places of interest (POIs), it could play a soundtrack which matches the visited POIs and enhances the experience. For example, a user visiting a Baroque cathedral might hear a classical composition by J.S. Bach, or during a visit to a lively market square the user might be offered an amusing folk tune.

In this paper we show that the user satisfaction for an itinerary can be increased by providing a musical soundtrack matching to the places of interest included in the itinerary. Furthermore, we show that simultaneously searching for pairs of matching POIs and music tracks can provide as good results as independently selecting a personalized set of POIs and then matching music tracks to the POIs. In other words, not only the music can be matched to a pre-selected itinerary, but also the discovered relations between music and a place can influence the itinerary selection process and, to a certain extent, substitute the POI personalization.

The main research challenge we have faced pertains with the complex relations between the two domains: locations and music. Some previous attempts of relating music to physical surroundings of the user include research on context-aware systems that use the information about the user's surroundings (time, weather etc.) to infer user's music preference [2], [3], [4]. Other works tried to project different domain representations to a common space, e.g. projecting text and music to a mood taxonomy. However, it is still unclear how (and which) properties of one domain relate to the properties of the other [5], [6]. Here, we have explored the combined usage of a limited number of well selected tags and textual descriptions as a way for matching items from different domains.

The rest of the paper is structured as follows. In order to give immediately a concrete example of the potential application area for the technology we are researching, we present the application used to conduct a case study (Section II). In Section III we describe the used music and POI matching techniques. In Section IV the evaluation procedure and results are presented, and finally in Section V we draw conclusions and point out future work directions.

II. Case Study Application

We start with presenting the architecture of the recommender system designed for evaluating possible matching techniques of music and POIs. Figure 1 shows the logical architecture of the system. The system uses profile information for standard item personalization as well as music-to-POI similarity for combining music and POI in a joined recommendation. The main components are: the user profile, the POI profile, the music profile and the recommendation algorithm that consists of POI ranking, music filtering and music-to-POI similarity computation. We will now describe these components in detail.

The user profile contains the basic music and sightseeing preferences of the user. The sightseeing preferences are used for ranking the POIs. The user preferences can be any combination of: art, architecture, history and nature. These four categories roughly correspond to the types of POIs stored in our database. The objects include castles, museums, monuments, churches and nature objects. Another part of the user profile contains the preferred music genre; our database contains music tracks belonging to classical and rock music

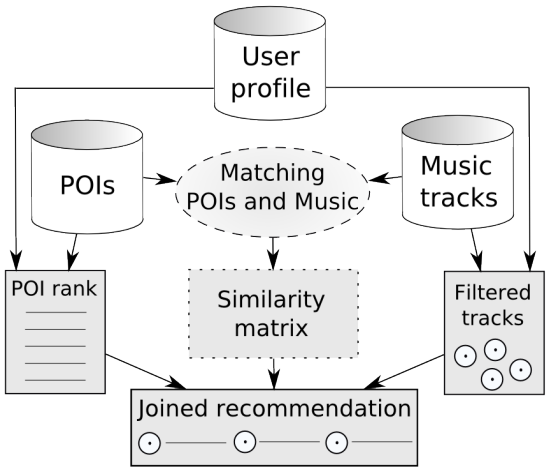


Fig. 1. Model of the system

genres. A user preferring classical (rock) music is offered only recommendations with soundtracks from the classical (rock) repertoire. In order to minimize the possibility that a user will not like a certain track, we selected some of the most famous compositions from both classical and rock music to be used in the system.

The music profile contains the title of the track, the description of the track, the genre of the track and a set of tags describing the music item. Table I shows an example of such a music profile. The music database, which was used in the experiments, consists of a set of manually selected music tracks (around 20 tracks per genre). The descriptions of classical music tracks have been taken from a dedicated web site (Classical Notes [7]), while the descriptions of rock music tracks were taken from Wikipedia [8]. The tags describing each music item have been assigned to them manually (see later).

Finally, the POIs are described by a POI profile containing: the name of the place of interest, the description the POI, a set of tags describing the POI, and the categories of the POI: art, architecture, history, nature. Table I gives an example of POI profile. 57 POIs in the city of Bolzano and its surroundings have been manually selected. The descriptions of the POIs have been taken from the local tourism web-sites [9], [10], [11].

The tags describing both music tracks and POIs were chosen from the "List of Adjectives in American English" [12] (see Table II) and assigned to them manually through a separate tagging application. 38 adjectives were selected in such a way that each adjective could describe both a location and a music track.

In order to efficiently tag all the items in our database we have implemented a web application that allowed the users to tag POIs and music tracks. The application was used by around 30 people. On average, 11 tags were assigned to each POI and 13 tags to each music track. Some additional information on the tags' statistics is given in Section IV.

TABLE I
EXAMPLE OF MUSIC AND POI PROFILES

Music profile	**POI profile**
Name: Wagner - Tristan und Isolde	*Name*: Roncolo Castle
Genre: Classical	*Type*: Art, Architecture, History
Description: Tristan und Isolde (Tristan and Isolde) is an opera in three acts by Richard Wagner to a German libretto by the composer, based largely on the romance by Gottfried von Strazburg...	*Description*: First mentioned in 1237, Runkelstein castle is situated in a picturesque position on a massive mound of porphyry rock at the entrance to the Sarntal valley. It has become famous for its superb secular fresco cycles...
Tags: Beautiful(2), Big(1), Bright(1), Calm(1), Clear(1), Cold(1), Colorful(1), Dark(2), Fast(1), Gentle(1),...	*Tags*: Beautiful(1), Big(2), Calm(1), Cold(1), Dark(1), Dull(1), Gentle(1), Happy(1), Mysterious(5), Narrow(1),...

TABLE II
THE SET OF TAGS FOR DESCRIBING MUSIC AND POI

Sad	Happy	Wide	Narrow
Scary	Amusing	Flat	Uneven
Angry	Calm	Dark	Bright
Embarrassing	Pleasant	Dull	Colorful
Mysterious	Clear	Powerful	Feeble
Heavy	Light	Ugly	Beautiful
Fast	Slow	Cold	Warm
Big	Small	Harsh	Gentle
Loud	Silent	Noisy	Melodic
Old	Modern		

The tags assigned to POIs and music tracks by the users have not been post-processed. As a result, profiles of some items may contain contradicting tags. In the example in Table I we see tags "Bright" and "Dark" assigned to the same item. This is natural, since different users often have different perception of the same music track or POI. The tagging process was conducted some months before the itinerary recommendation experiment. Each user was offered a randomly selected POI item from our database of POIs. After tagging the POI, the user was given a randomly selected music track from or database which was followed by another random POI. The sequence continued until the user decided to logout or tagged all the items in our database. The users could choose the genre of the music tracks they want to tag. This was done to avoid forcing the users to tag music they do not like.

III. THE RECOMMENDATION PROCESS

In this work we evaluated three approaches for combining music and POIs that will be describe later. Here we start with describing music filtering and POI ranking since they are used in all the approaches. The goal of both music filtering and POI ranking steps is to reduce (filter) the set of POIs and music tracks before the two types of items can be paired by one of the POI and music recommendation algorithms. In fact, in all the three algorithms the recommendation process starts with the input of the user's ephemeral preferences, which are described by the tags entered by the user. Based on these ephemeral

preferences and the user's profile data the ranking of the POIs and the filtering of music tracks is initially done separately.

Since we are not interested in the best quality single-domain recommendations, the music tracks are simply filtered by the user's preferred genre. POI ranking is more complex, and consists of measuring the overlap between the categories of POI and the user's interests as well as the tags entered by the user and the tags stored in the POI profile. The POI ranking algorithm takes as an input the user's profile data, the user's ephemeral preferences, the set of all POIs and the number of POIs to return (N). The result of the algorithm is the top N ranked POIs. The user's sightseeing preferences are defined in the user profile (a subset of {art, architecture, history, nature}) and the user's ephemeral preferences, $userTags$, are given as input by the user when requesting an itinerary recommendation. Only POIs that have at least one common category with the user's sightseeing preferences are considered for recommendation. The tags describing a POI, $POItags$, is a bag of words where tags may be repeated (if assigned to a POI by different users). The algorithm then computes the intersection of the bags $userTags$ and $POItags$. The cardinality of the intersection is stored as the score of POI.

Note that we take into account the repeating tags in POI profile, i.e. an intersection of the POI tags {sad, sad, sad, cold} with user's ephemeral preferences {sad, cold, dark} would give an intersection of 4 elements: {sad, sad, sad, cold}. This is done under the assumption that repeating tags are more important since more than one user associated a POI with these tags. Therefore, the presence of such important tags should be reflected in the POI score. After looping through all the POIs in the database, the algorithm sorts the ranked list according to the score of POIs and returns the top N POIs.

The outputs of both music filtering and POI ranking processes, i.e., a set of music tracks and a ranked list of POIs, are then combined in three possible ways:

- **POI rank.** In this approach, we assign to each of the top N POIs (from the ranked list of POIs) one music track chosen randomly from the set of filtered tracks. In this way the selected music is liked by the user but is not adapted to the POI. The order of the recommended POIs is determined by POI rank.
- **POI rank + music matching.** In this approach, after obtaining the ranked list of POIs, we assign the best-matching music track (from the filtered tracks) to each of the top N POIs. The music-to-POI similarity computation is described in Section III-A. Here, the order of recommended POIs remains identical to the first approach. Only the recommended music is different.
- **Music similarity rank.** In this approach, we do not simply assign music to the existing ranking of POIs, but rather combine the POI rank with the music-to-POI similarity score to produce a new ranking for the pairs of POIs and music tracks. We then recommend the top N pairs. The POI-music pair score is computed using the following formula:

$$score = (0.7 * POIpers) + (0.3 * similarity) \quad (1)$$

Here $POIpers$ is the POI score obtained from the POI ranking algorithm and $similarity$ is the music-to-POI similarity (see Section III-A).

The weights in formula 1 have been set after running experiments with different parameter values and analyzing the obtained itinerary recommendations. The aim of parameter tuning was to obtain itineraries that still reflected the user's ephemeral preferences (represented by $POIpers$), but also contain some of the highly matching music-POI pairs (represented by $similarity$). As assigning more weight to the $similarity$ parameter resulted in recommendations that did not reflect the user's preferences for the POIs, more weight was gradually shifted towards $POIpers$.

A. Computing Music-to-POI Similarity

The music-to-POI similarity scores are stored in a matrix that is computed off-line since it depends only on static item profiles. The similarity matrix S is a P x M matrix where P is the number of POIs in the database and M is the number of music tracks in the database. Each element of the matrix S_{ij} represents the similarity of music track j to POI i . Each similarity value is computed based on the profiles of the POI and the music track as follows:

$$similarity = (0.2 * tagOverlap) + (0.8 * textSc) \quad (2)$$

where $tagOverlap$ is the cardinality of the intersection of the tags in the POI profile with the tags in the music profile and $textSc$ is computed as follows:

$$textSc = \begin{cases} ldaSc, & \text{if } tfSc < 0.1 \\ tfSc, & \text{if } tfSc \geqslant 0.1 \end{cases} \quad (3)$$

In the above formula $ldaSc$ is the similarity of text descriptions in item profiles based on Latent Dirichlet Allocation *LDA* [13]. Using *LDA* we represented each document as a 100-dimensional vector of topic probabilities $d = (p_1, p_2, ..., p_{100})$, where p_i is the probability that the i^{th} topic is found in document d ($p_i \in [0, 1]$). The similarity between two text documents is then computed as the cosine distance between the two document vectors $d_1 = (p_1^1, ..., p_{100}^1)$ and $d_2 = (p_1^2, ..., p_{100}^2)$.

Likewise, $tfSc$ is the text similarity of the item profiles texts computed using classical *TF-IDF* representations [14]. Each document is represented by a term vector $d = (t_1, t_2, ..., t_n)$, where t_i is the weight of the i^{th} term in the document d and n is the number of distinct significant terms found in the document corpus (5595 in our case). The similarity between two text documents is then computed as the cosine distance between the two vectors $d_1 = (t_1^1, ..., t_n^1)$ and $d_2 = (t_1^2, ..., t_n^2)$.

Experiments have shown that the two text processing algorithms (*LDA* and *TF-IDF*) produce different results depending on the type of text documents. *TF-IDF* can give good results

when two documents are linked by highly weighted keywords. However, such situation is not common in the dataset used in this research due to different nature of the described objects. POIs and music descriptions typically lack direct links between themselves. In such cases *LDA* algorithm performs better than *TF-IDF*. The $tfSc$ threshold in formula 3 was therefore set by comparing *TF-IDF* similarity values against human judgment of document relevance for our data corpus.

The weights in formula 2 as well as the number of topics in *LDA* computation have been established empirically. Due to the high cost of live-users experiments these parameters have not been optimized. More extensive testing and parameter tuning is a topic of future work.

IV. Evaluation and Results

In order to validate our approach we implemented a web application that offered the user itinerary recommendations for the city of Bolzano and its surroundings. An itinerary was composed of three places of interest inside and around the city (monuments, buildings, nature parks etc.) and a soundtrack (one music track per POI).

The human/computer interaction with the evaluation application consisted of three main parts: filling out the profile information (registration), viewing two recommended itineraries and providing a feedback, i.e., choosing the preferred itinerary.

Before using the system the users registered providing their age, gender, country, preferred music genre, sightseeing interests and the level of knowledge about Bolzano city. In the starting page, the user was asked to enter the ephemeral sightseeing preferences - a few tags describing the POIs that she would like to see. The tags were chosen from a finite set of tags that was previously used to tag both POIs and music tracks. After specifying the preferences for POIs, the user was given a first recommended itinerary. The suggested sequence of POIs was displayed, one POI after the other, and the accompanying music tracks were played. Having viewed the first itinerary, the user was offered a second itinerary that satisfied the same ephemeral preferences but was computed with a different version of the recommendation algorithm (see Section III).

Finally, after viewing the two itineraries, the user was asked to choose the itinerary that she liked the best. The users were not aware of the different modes used in the recommendation process. Therefore, recording the user choices and analyzing the data allowed us to compare the user satisfaction with different recommendation techniques and to analyze the effect of music matching on the quality of recommendations.

The users participated in the evaluation on a voluntary basis. Invitations to participate have been sent to research colleagues as well as readers of the user-modeling mailing list (um@unito.it). A total of 53 users used the system. 13 users used the system 3 times and more. 40 users used the system once or twice. In total, 115 evaluation sessions were performed.

In our study we compared the number of times each recommendation algorithm was preferred by the users in the evaluations. Figure 2 shows the selection probability of the three algorithms considering all session data together, and separately sessions where classical or rock music were the preferred genre of the user. Selection probability is estimated as the ratio of times an itinerary with a soundtrack generated by one method was selected over the total number of times it was offered in one of the two suggested itineraries in a recommendation session. In this figure, *POI rank*, *POI rank + matching*, and *music similarity rank* are the different versions of recommendation algorithm.

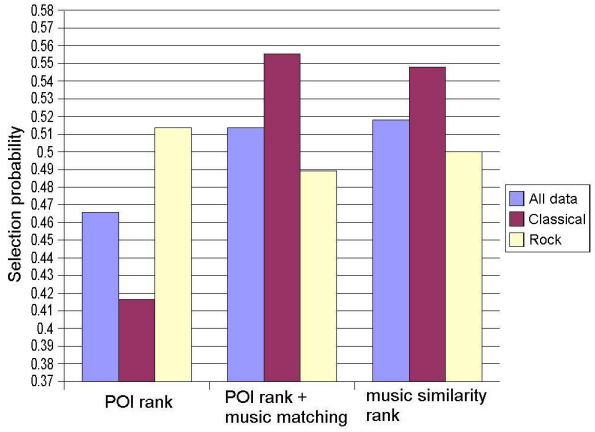


Fig. 2. Selection probability of different algorithms

Considering all session data we see that both *POI rank + matching* and *music similarity rank* approaches are preferred more often than the recommendations where music is assigned to POI without matching. This data supports our main hypothesis that the process of music matching to POI is beneficial and produces more satisfying recommendations. We note that even in the simpler algorithm, *POI rank*, the played music was selected among tracks of the user's preferred genre. So the played music was among the preferred type in all the three cases.

The results do not show a clear preference of the users for one of the two approaches: *music similarity rank* and *POI rank + matching*. In fact, these approaches have similar selection frequency. While in the *POI rank + matching* approach the recommended itineraries are primarily composed by ranking the POIs and only then assigning music track to each POI, the *music similarity rank* approach focuses rather on recommending pairs where POI and music are highly matching to each other. Therefore, we conclude that selecting highly matching pairs of POI and music may, to some extent, compensate the lack of a precise POI personalization.

The results for session data collected separately for each music genre show that while classical music listeners clearly preferred the advanced recommendation techniques, rock music lovers did not show any preference among the different approaches. This could mean that even the simpler approach, i.e., assigning a generic track in the user's preferred genre can be suitable for certain music genres. The results suggest

that the effectiveness of the two proposed approaches for matching music with POI depends on the genre of music preferred by the system users. Genre dependencies in music information retrieval have been identified in previous works. For instance, when classifying music into emotional categories Li and Ogihara [15] obtained better results within certain music genres (classical and fusion music).

The dependence of the algorithm performance on music genre could be explained by the fact that classical music is much more diverse (in tone, melody, rhythm, mood and tempo changes) compared to rock music. Therefore, classical tracks can be easier for the users to tag with a wider variety of tags which makes the matching procedure more effective. This is backed up by the fact that out of more than 2000 possible pairs of music and POIs in our database the top 20 pairs contain only classical music. In order to investigate the possible reasons for this, we have checked the tag distribution for music items in our database. Figure 3 shows the distribution of tags assigned to music of different genres by the users. We computed the probability of each tag usage as $\frac{tags}{totalTags}$, where $tags$ is the number of times a tag was assigned to an item and $totalTags$ is the total number of tags assigned to the items in a collection (either classical or rock). The figure 3 shows that the tag distribution of the classical music tracks is more uniform (flat). In other words, the users used more diverse tags while tagging classical music tracks.

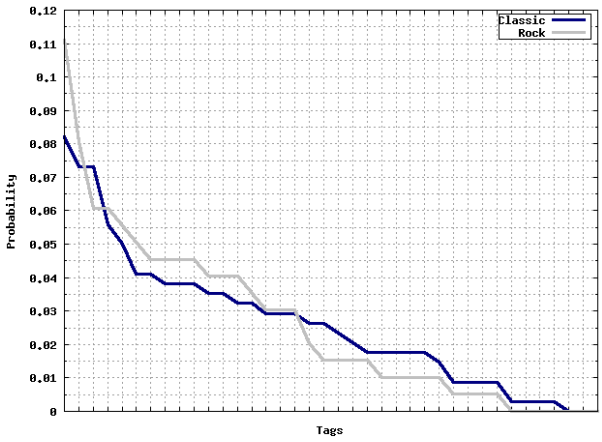


Fig. 3. Tags usage probability among the different music genres

Interestingly, during the tagging process rock music received less tags compared to classical music tracks (although similar number of users were tagging both types of music). We have collected on average 16.24 tags for each classical track and 9.9 tags for each rock track in our database. Moreover, during a single tagging session, a user tagging rock music assigned to a track an average of 4.3 tags. Meanwhile, a classical music listener assigned to a track an average of 5.68 tags. Conversely, we have collected on average 11.18 tags for each POI in our database. A user tagging POIs assigned an average of 3.79 tags to a POI.

We have also checked which types of tags were used when tagging POIs and music. We have identified 3 general categories for the 38 adjectives used in the tagging process: emotional, sound quality and physical quality adjectives. We compared the frequency of tags of each type in POIs and music respectively. The frequency of each tag type was computed as a ratio of tags of this type in POI/music profiles over the total number of tags of this type in our database. Figure 4 shows that POIs are mostly described by physical quality adjectives while music is most often described by sound quality adjectives, as expected. Conversely, emotional tags are almost equally distributed between POIs and music. Therefore, we conjecture that emotional tags could allow a more efficient music-to-POI similarity computation. This is an hypothesis that we want to test in the future.

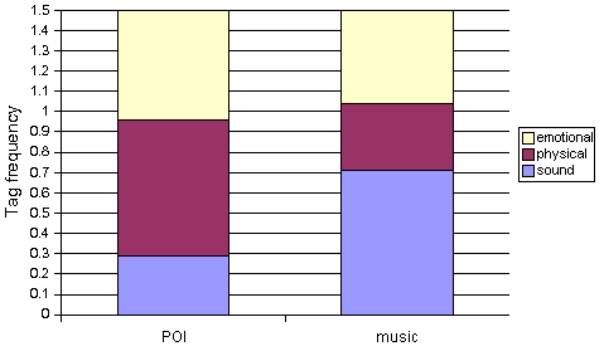


Fig. 4. Tags usage frequency in POIs and music

In fact, there have been already some attempts to use emotional labels for music classification [16]. Several works have studied and selected a restricted set of emotional labels that can describe music. Hevner [17] suggested 8 clusters of descriptive adjectives that were revised and expanded by Farnsworth [18]. A more recent and elaborated research on emotions evoked by music was carried out by Zentner et al. [19]. The proposed Geneva Emotional Music Scale (GEMS) consists of 9 emotional clusters with 4-5 adjectives in each group. We plan to use this model for re-tagging the items before further experiments.

V. Conclusions and Future Work

In this paper we have presented a matching technology for building multimedia and cross-domain recommender systems. We have validated two hypotheses: a) user satisfaction for a POIs recommendations can be increased by accompanying POIs with music tracks that are matching the user's profile and the POIs; and b) simultaneously selecting best-matching pairs of POIs and music tracks can provide as satisfying recommendations as those computed when first personalized POIs recommendations are selected and then the music is matched to each of these POIs. In fact, the general result of our live users experiments is that users more often prefer the itineraries with soundtrack music matching to POIs.

In particular, we have shown that the approach based on simultaneously selecting music and POI provides similar results to the approach where the music is matched to pre-selected

POIs. This could mean that the high similarity between a music track and a POI can compensate the lack of POI personalization. However, such results can also be caused by the relatively small set of music tracks, which resulted in similar music tracks recommended in the two approaches. Hence, this is an issue that deserves some more study and experimental tests.

Furthermore, the evaluation results showed that the performance of the implemented algorithms depends on the music genre preferred by the users. Our system offered the users two musical genres: classical and rock music. Recommendation sessions where classical music was chosen tend to give better results with the two music matching techniques that we proposed. While such results could have been determined by the relatively small size of the used data set, we believe that this is influenced by the nature of music genres; rock music being more homogeneous and difficult to tag.

This research is part of an ongoing work on tag-based similarity techniques for heterogeneous objects. Future work includes fixing some of the limitations of the current approach and performing a more extensive user study with more users, revised POIs and more music tracks. We observe that user studies in recommender systems are rare and most of the validations have been performed off-line; cross validating the accuracy of the recommendation in a given data set [20], [21]. So, our work is still among the few attempts to evaluate the recommendations with real users' feedback rather than trying to reproduce the past user evaluations on recommended items.

Before conducting new user studies we intend to revise the used data set. We will focus on classical music, since rock music did not perform well in the experiment described in this paper. Another potentially useful music genre could be movie soundtracks, since these tracks typically carry a strong emotional message and could be easier to tag. We also intend to group the POIs in a small number of categories, and to tag the categories rather than the instances. This should make the categories of POIs more unique and distinctive, and we conjecture that different categories will receive different tags. Most importantly, we are revising the set of adjectives used for tagging music and POIs. We are considering using the emotional adjectives from the GEMS model [19].

With respect to the algorithms used in the proposed approach, we note that the computation of the matching of two items as the intersection of the tags assigned to them should perform better when it is normalized by the number of tags present in the items' profiles. Furthermore, it is important to perform additional testing (cross validation) for tuning the different weights used in the matching formulas. We are also considering some newly proposed similarity functions for tagged objects that have been already validated in other applications [22].

We conclude by observing that the described technique could be used in many applications that requires the personalized presentation of POIs, e.g. in tourism web-sites, and electronic guides. They can also be applied in mobile applications for personalizing the travel experience with ad hoc music adapted to the location of the user. In particular, we believe that the proposed techniques can be very useful to filter and personalize stream audio content.

REFERENCES

[1] S. Berkovsky, T. Kuflik, and F. Ricci, "Mediation of user models for enhanced personalization in recommender systems," *User Modeling and User-Adapted Interaction*, vol. 18, no. 3, pp. 245–286, 2008.

[2] H.-S. Park, J.-O. Yoo, and S.-B. Cho, "S.b.: A context-aware music recommendation system using fuzzy bayesian networks with utility theory," in *FSKD 2006. LNCS (LNAI)*. Springer, 2006, pp. 970–979.

[3] A. Ranganathan and R. H. Campbell, "A middleware for context-aware agents in ubiquitous computing environments," in *Middleware*, ser. Lecture Notes in Computer Science, M. Endler and D. C. Schmidt, Eds., vol. 2672. Springer, 2003, pp. 143–161.

[4] S. Reddy and J. Mascia, "Lifetrak: music in tune with your life," in *HCM '06: Proceedings of the 1st ACM international workshop on Human-centered multimedia*. New York, NY, USA: ACM, 2006, pp. 25–34.

[5] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[6] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma, "Musicsense: contextual music recommendation using emotional allocation modeling," in *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*. New York, NY, USA: ACM, 2007, pp. 553–556.

[7] P. Gutmann, "Classical notes," http://www.classicalnotes.net/.

[8] "Wikipedia," http://en.wikipedia.org/.

[9] "The official website of south tyrol," http://www.suedtirol.info/.

[10] "Eppan tourism website," http://www.eppan.com/.

[11] "Ritten tourism website," http://www.ritten.com/.

[12] P. Noll and B. Noll, "List of adjectives in american english," http://www.paulnoll.com/Books/Clear-English/English-adjectives-1.html.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[14] G. Salton, *Automatic Text Porcessing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley, 1989.

[15] T. Li and M. Ogihara, "Detecting emotion in music," in *Proceedings of the fourth international conference on music information retrieval (ISMIR, 2003)*, Baltimore, USA, October 2003. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.8122

[16] D. Baum and A. Rauber, "Emotional descriptors for map-based access to music libraries," in *Proceedings of the 9th International Conference on Asian Digital Libraries*, November 2006, pp. 370–379.

[17] K. Hevner, "Experimental studies of the elements of expression in music," *The American Journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936. [Online]. Available: http://www.jstor.org/stable/1415746

[18] P. R. Farnsworth, *The social psychology of music*. Iowa State University Press, 1958.

[19] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, August 2008. [Online]. Available: http://dx.doi.org/10.1037/1528-3542.8.4.494

[20] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[21] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transaction on Information Systems*, vol. 22, no. 1, pp. 5–53, January 2004.

[22] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme, "Evaluating similarity measures for emergent semantics of social tagging," in *WWW '09: Proc. of the 18th international conference on World wide web*. ACM, 2009, pp. 641–650.