AUGMENTING VIRTUAL-REALITY ENVIRONMENTS WITH SOCIAL-SIGNAL BASED MUSIC CONTENT

[1]Ioannis Karydis, [2]Ioannis Deliyannis and [2]Andreas Floros
{[1]Department of Informatics, [2]Department of Audio and Visual Arts} Ionian University, Corfu, Greece.

ABSTRACT

Virtual environments and computer games incorporate music in order to enrich the audiovisual experience and further immerse users. Selecting musical content during design-time can have a controversial result based on the preferences of the users involved, while limiting the interactivity of the environment, affecting thus the effectiveness of immersion. In this work, we introduce a framework for the selection and incorporation of user preferable musical data into interactive virtual environments and games. The framework designates guidelines for both design and run-time annotation of scenes. Consequently, personal music preferences collected through local repositories or social networks can be processed, analysed, categorised and prepared for direct incorporation into virtual environments. This permits automated audio selection based on scene characteristics and scene characters' interaction, enriching or replacing the default designer choices. Proof-of-concept is given via development of a web-service that provides a video game with a dynamic interactive audio content based on predefined video game scene annotation and user musical preferences recorded in social network services.

Index Terms— social-signal processing, music information retrieval, social-networking, user immersion

1. INTRODUCTION

Virtual-reality (VR) aims in producing synthetic environments wherein users experience a subjective sense of presence [1]. While most relevant research focuses on graphics and interaction related topics, the importance of the audio within VR has recently attracted increased attention [1-5], lending mainly from the fields of film sound/music and auditory display studies.

The importance of auditory stimulus and especially musical content within VR worlds can easily be demonstrated using video games as an example. Therein, audio functions as a means to contribute to usability, mood and atmosphere, orientation and control of the fictional environment, supporting thus the sense of presence in the VR world [3,5]. In addition, the seemingly meaningless or rather difficult to further elaborate on, importance of audio is becoming apparent to players when muted [2, 6-8].

Contrasting the auditory stimulus requirements in media such as movies [9, 10] and theatre [11] to those in interactive media including computer games, art-performances and VR worlds [7], it becomes clear that interactivity introduces complexities during their development in terms of appropriate non-linear audio composition and editing for use within the application scenario. These are commonly addressed during application design time and in most cases are compilations from a fixed set of audio effects and music for each scenario within the art-performance or application. Recent research on different types of video game sound argues the use of dynamic audio; that is, audio that reacts to changes in the game-play environment or in response to a user event [12]. Even in the latter cases, the breadth of customisation remains within design time decisions.

In order to make the virtual reality experience more customised and personal, aiming in deeper immersion, the use of appropriately selected musical content may be of great importance. As VR users' aesthetic experience is argued [13] to be pertinent to the musical content of the environment, each user's musical choices appear as natural candidates for the musical decoration of the environment. Nevertheless, for audio to retain its functional aspects within the VR environment [3], methods are required for the selection of audio from the user's choices based on design-time specifications, run-time scene characteristics and run-time characters' interaction. Music Information Retrieval (MIR) offers such methods for the management of musical material. In addition, MIR methods have recently [14] extended to harness the high importance contextual information contained in tags assigned on musical content in Web 2.0 services.

While design time annotation by VR environment designers requires a shift of paradigm in such environments production, run time annotation can heavily rely on Social Signal Processing (SSP) [15] as well as Affective Computing (AC) [16] methods.

1.1. Motivation

This research focuses on the enhancement of users' immersion within virtual reality environments by means of the auditory stimulus. In such environments, research on audio functionality shows that the role of audio is far beyond simply entertaining and in some cases is reported as "overarching" [3]. Nevertheless, the current VR audio

978-1-4577-0274-7/11/$26.00 ©2011 IEEE DSP2011

environment design paradigm does not consider the use of user selected audio following designer's principles. In some cases user selected audio is an option, but the auditory characteristics of environment's designer original aim are not taken into consideration. In addition, the current practice of design-time auditory annotation of scenes can be restrictive for the breadth of interactivity of a scene, neglecting run-time affective state and social interaction between users and environment's characters as well as scene characteristics. Even in cases of real-time music generation for such dynamic environments, despite research indications [13], user preferable audio content is not considered.

1.2. Contribution and Paper Organisation

Herein, we propose a framework for the incorporation of user preferable musical content into interactive virtual reality environments based on generic designer scene annotation of the VR environment during design-time as well as run-time annotation based on scene characteristics and user - scene characters' interaction. Accordingly, for the aforementioned motivating requirements, our contributions are summarised as follows:

We propose a generic framework for the incorporation of users' musical preferences within a VR environment following the design-time annotation in order for audio to retain its original functionality. The framework also proposes the use of the social signals during scene users' interaction that convey important contextual information for the automated run-time annotation of a VR environment scene and thus its appropriate musical decoration.

We present a proof-of-concept implementation of the proposed framework in a video game utilising user preferable musical content posted in social network services in order to dynamically retrieve and incorporate audio in run-time, following design-time game scene annotation.

The rest of the paper is organised as follows. Section 2 reviews related work and Section 3 presents the proposed framework for incorporation of user music into VR environments. Section 4 presents a proof-of-concept implementation of the proposed framework in a video game, while the paper is concluded in Section 5.

2. BACKGROUND AND RELATED WORK

Recent developments of audio technology for video games are briefly presented together with relevant research results on music information research, while an overview of social signal processing & affective computing issues is also presented.

2.1. Game Audio

Initial attempts to video games date back to 1958 with William Higginbotham's "Tennis for Two" and "Spacewar!" (1962) that had no sound at all. The first game to include a soundtrack was Taito/Midway's "Space Invaders" (1978), featuring a continuously looping background sound. In addition, the soundtrack had also a dynamic character, as it would increase tempo for increasing game progress.

Development of game audio advanced slowly based on the capabilities of the available platforms until 1983 that the MIDI protocol was defined. MIDI allowed predefined musical devices to be utilised by solely issuing commands, avoiding thus the incorporation of actual sounds leading to small sound file sizes in addition of the ease of usage due to standardisation of the protocol. The widespread penetration of the CD-ROM technology led to recorded music availability for video games, in addition to more memory to store better graphics. However, CD-ROM also provoked the abandonment of MIDI audio for games and thus the notion of dynamic musical content. The latest generation of games focuses on recorded audio, in many cases using full orchestras and choirs, 3D and multi-channel surround sound, DVD-quality. Nonetheless, despite having increased processing and storage capabilities, key aspects of dynamic game sound are yet to be explored [6].

2.2. Music Information Research

Following the definition provided by Futrelle and Downie [18], music information retrieval research is a rapidly growing interdisciplinary area encompassing computer science, information retrieval, musicology and music theory, audio engineering and digital signal processing, cognitive science, library science, publishing and law. Its agenda mainly pertains to the development of methods for the management of musical material for retrieval, classification, analysis, transcription and recommendation, among other prominent uses.

The establishment of Web 2.0 has offered an initial spark as well as a unique opportunity for the requirement of MIR concerning the exploitation of contextual knowledge for its methods. The effect of contextual information can easily be illustrated considering one of the principal MIR processes, the assessment of musical similarity. Measuring the similarity between two musical pieces is widely accepted to be a hard problem, as it is difficult to be defined in strictly objective terms [19]. Despite the inherent difficulties in assessing music similarity, its output is of great importance to numerous areas of MIR.

Initial methods computed music similarity using objective metadata, e.g., composer name, song title, etc. However, such methods are in cases not as effective as content-based methodologies, since metadata usage requires prior knowledge of data that is not conveyed by listening, may be unavailable and have limited scope due to usage of pre-defined descriptors. Content-based similarity has been under extensive research [20-23] focusing on features extracted from the audio content, which express different attributes of a musical datum. Nevertheless, it has been shown that the performance of content-based music

similarity is reaching a limit, characterised as "glass ceiling" [20], which is far from the best achievable. Bearing in mind the aforementioned subjectivity of musical similarity and the nature of user assigned tags on musical data, it comes as no surprise that methods measuring musical similarity based on tags are very frequently more accurate than content-based methods [24]. Nevertheless, rich as it may be in contextual characteristics, the information provided by Web 2.0 social services is known to present disadvantages [14].

2.3. Social Signal Processing for MIR

Social signal processing [25] is the innovative and multidisciplinary research and technological domain that aims at providing computers with the ability to sense and understand human social signals. Despite being in its initial phase, SSP has already attracted the attention of the technological community. Social interactions between humans or between humans and computers offer a wealth of non-verbal communication which convey information about mental state, personality, and other traits of people, offering thus a great deal of contextual information for the purposes of MIR methods [15].

During social interactions, non-verbal behaviour not only conveys this information for each of the involved individuals, but also determines the nature and quality of the social relationships they have with others. This happens through a wide spectrum of non-verbal behavioural cues that are perceived and displayed mostly unconsciously while producing social awareness, i.e., a spontaneous understanding of social situations that does not require attention or reasoning.

Following the taxonomy of SSP, the key behavioural cues occurring during social interaction include: (i) physical appearance, (ii) gesture and posture, (iii) face and eye behaviour, (iv) non-verbal vocal behaviour, (v) non-linguistic vocalisations and (vi) space and environment. All the aforementioned cues offer important information for the purposes of MIR and VR environment annotation that are still very little explored in the literature.

2.4. Affective Computing for MIR

Emotions have a major impact on essential cognitive processes and play a dominant role in human creativity, intelligence, human thinking and decision-making [16]. In order to develop computers that interact naturally and intelligently with humans the ability to at least recognise and express affect is necessary. Accordingly, a variety of available physiological measurements yield clues to one's hidden affective state far beyond the previously mentioned social behaviours.

Affective computing's key process of "recognising emotions" refers to measuring observations of human motor system behaviour and its respective VR character that correspond with high probability to an underlying emotion or combination of emotions. Such measuring of observations is nowadays hugely supported by a number of currently existing technological fields in the real world, while in VR environments such measurements are far easier to be performed. Once again, the assessment of the affective state of both the user and/or the VR character offer important information for the purposes of MIR and VR environment annotation that are still very little explored in the literature [26].

TABLE I
WEB-SERVICE STEPS FOR USER'S MUSICAL PREFERENCES SPECIFICATION AND CONTENT QUERY/DELIVERY USING LOCAL CONTENT AND SOCIAL NETWORK CONTENT

Step	Local Content	Social Network Content
1	User specifies local content files	User specifies social network credentials to get favourite content
2	VR system sends metadata of selected content to web-service and requests content	VR system requests content
3	Web-service identifies content	Web-service identifies content
4	Web-service returns content query reply	Web-service returns content query reply
5	----	Content cached locally or streamed

3. PROPOSED FRAMEWORK

The initial consideration of the proposed framework is in the audio design paradigm shift required during design-time in order to allow for customised musical content in VR environments. In addition, as previously mentioned, run-time social interaction between users and/or characters of the VR environment may have an impact on the selection of the musical content and thus must be addressed too. Finally, the proposed framework deals with the MIR web-service furnishing information over user preferred content as well as access to freely available musical content.

3.1. VR Environment Audio Design Shift

In order to incorporate audio content that is unknown during design-time in VR environments and at the same time retain the designer's aim as to the functionality of the experience, the proposed framework identifies the following requirements:

--at design time, scene annotation is required using objective metadata (e.g. beats-per-minute, genre) of the content required to sound,

--at design time, scene annotation is required using descriptive tags (e.g. dramatic - instrumental - epic, mellow - love - jazz) of the content required to sound,

--at run time, using the Social Signal Processing module, scene annotation is required based on the social signals of social interactions between users and/or scene - environment

characters' interaction revealing information about feelings, mental state, personality, and other traits that provide a great deal of contextual information and can be used MIR search as tags,
--at run time, using the Affective Computing module, scene annotation is required based on “recognised emotions” from observations of user or VR characters motor system that correspond to underlying emotions offering high contextual information to be used in MIR search as tags, and
--at design time, specific content selected by the designer, fitting for the scene, needs to be designated and made available in order to allow MIR methods to identify similar content using content-based similarity as well as have an effective alternative in case the user's preferred content proves unsuitable.

3.2. Run-time SSP & AC Modules

In order to fully customise the VR experience taking advantage of the social interaction between characters within the environment at run-time may be of great importance. Moreover, corresponding actions of characters, based on probability, to an underlying emotion or combination of emotions offers additional important information. Both types of information may assist in the attempt to identify plausible selections to the auditory stimuli of the environment and thus increase immersion further. Both SSP and AC fields are still in infancy and therefore no concrete results exist. The authors believe that the development and utilisation of such modules, incorporated within the logic of the VR environment, will boost the dynamic character of the environment, enabling the development of highly adaptive virtual reality systems and computer game environments.

3.3. Web MIR for VR Environment Service

To begin with, the user's musical preferences need to be identified. Herein, we assume that such preferences can be gathered through social networking or by purchased locally available content, without losing generality as any alternative input source can be utilised. Accordingly, in case of the later, the user must provide credentials for the respective social network account. Then, the VR environment requests content based on design-time specified needs, such as tags, metadata, and similarity to designer's choices or extracted features. After selection of the appropriate content from the user's musical preferences, the web service returns content pointers in case of local content or caching URIs of the reply content, in case of user musical preferences from social network. Table I shows the steps required to obtain content. In the case of a run-time content request, steps 2, 3 and 4 are required, while if the user content is not local, streaming methodologies can be utilised in order to ensure timely acquisition of musical content. Appropriate content identification based on the VR environment's request can be implemented using all three types of information - metadata, content extracted features and contextual information - in order to perform similarity measurement, as described in Section 2.2.

4. PROOF-OF-CONCEPT

In order to provide proof-of-concept for the proposed framework, we have implemented a partial functionality Web MIR for VR Environment Service (Web MIRVRES) and adapted an open version of the Arkanoid video game [27], developed under the XNA programming environment [28] using c# programming language.

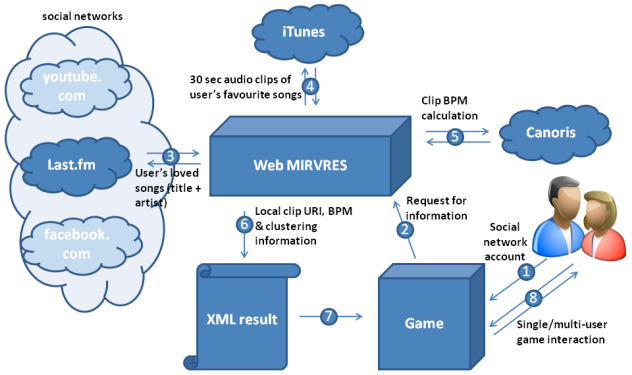

Fig. 1. Interaction and data exchange between the game and the Web MIRVRES

In order to demonstrate the interactive audio selection abilities, the game is adapted to utilise a game speed based routine for dynamic audio clip selection. Following the aforementioned framework, social signals in the form of user tracks tagged as "loved tracks" were extracted directly from the Last.fm user's account, using the service's API [29] and then embedded into the game during run-time. In addition, for the purposes of this demonstration, we have selected the utilisation of the Beats Per Minute (BPM) information of the audio content as objective metadata used to match equivalent design-time scene annotation. Figure 1 presents a schematic overview of the interaction and data exchange between the game and the Web MIRVRES, for the game to receive the user preferred musical content as well as accompanying information on each song.

4.1. Web MIRVRES and Querying Interface

The implemented Web MIRVRES centrally controls the overall process, sectioned in the following discrete steps:
--After receiving an Last.fm username the process initiates by retrieving the list of tracks the user has tagged as "loved tracks" from the selected social network service,
--The iTunes API is used to match the track and artist name received from Last.fm for each "loved track" of the previous step and download the respective audio clip samples,

--The samples are then transferred to Canoris service in order to detect the BPM for each track, following research by Gouyon [31] and Davies & Plumbley [32],
--The cached audio clips in the Web MIRVRES and the collected BPM information are then accessible via the resulting XML file that is passed to the information requesting application.

The querying interface is implemented in PHP scripting language and query arguments are passed through http GET requests to the Web MIRVRES's base-URI. The parameters implemented for this case-study are:
--maxNoOfSongs: the maximum integer value of returned songs (depending on the number of songs the user has tagged as "loved" in Last.fm account and their respective audio clip availability on iTunes), default value = 30,
--echoOn: a boolean variable that allows returning results in human readable version (if set to "true" no xml result is returned), default value = false,
--lastfm_username: a string that contains the Last.fm account username,
--noOfClusters: the (integer) number of clusters the songs identified are assigned to based on their BPM value, default value = 3

The results are returned using an XML structure that has one "Track" tag for each of the result tracks and for each such tag, the following nested tags: track_name; track_artist_name; bpm; uri; filename; cluster; (clustering using the k-means algorithm [30]).

4.2. Game Adaptation

The XNA Arkanoid game utilised for the experimentation has been adapted in order to change the audio sounding at any moment on account of the speed of the game. Starting the game leads to an intro where until the user decides to begin a new gaming session; the preferred tracks are played in random order. After initiating a game session, as the score of each stage increases, so does the ball's speed. For every 10.000 score points, the speed increases by a unit. At any stage of the game there are at maximum 8 different ball-speed levels, due to the number of tiles and thus the user's preferred tracks are categorised in 8 clusters based on their BPM using the "noOfClusters" parameter while querying the Web MIRVRES for content. Each stage has a maximum score and a maximum value of ball speed changes. Consequently, at each stage, according to the current level of speed of the ball and the maximum speed level possible for the stage, tracks from the respective BPM-cluster are reproduced, with higher BPM tracks selected for higher ball speeds. Figure 2 displays a screenshot of the XNA Arkanoid at the beginning of the first stage, where the ball speed is at the lowest possible. Consequently, the track sounding is selected from the cluster with the smaller BPM. Figure 3 shows a screenshot towards the end of the second stage, where a track from the fourth cluster is selected since the second stage began in 55000 score points. Figure 4 depicts the distribution of the songs used in the demo run.

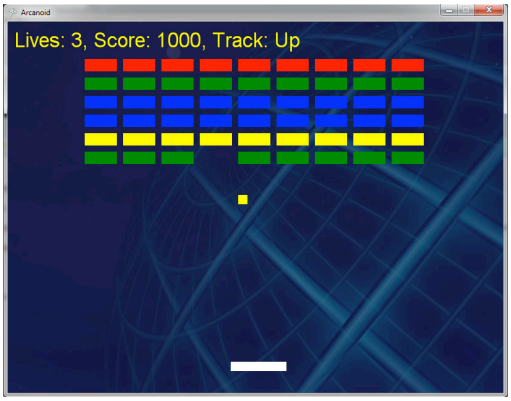

Fig. 2. The XNA Arkanoid game at the beginning of 1st stage, reproducing the 70BPM track entitled “Up” by Rob Crow.

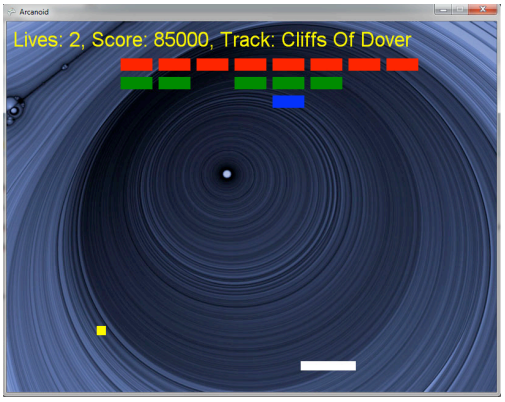

Fig. 3. The XNA Arkanoid game towards the end of 2nd stage, reproducing the 95BPM track entitled “Cliffs of Dover” by Eric Johnson.

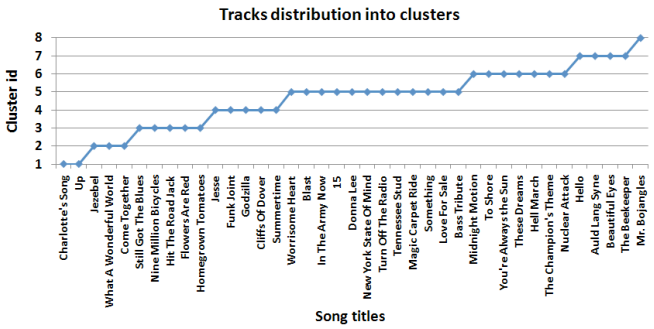

Fig. 4. The distribution of the songs into clusters for the demonstration execution.

5. CONCLUSION

In this paper, we presented the practical utilisation of a framework for enhancement of users’ immersion within virtual reality environments, which enriches existing systems with advanced auditory stimulus functionalities. Various developmental issues were discussed under a real-life case scenario. The proposed framework describes the incorporation of user-preferable musical content into

interactive virtual reality environments based on generic designer scene annotation of the VR environment during design-time as well as run-time annotation based on scene characteristics and user - scene characters' social signals and affective state.

In order to provide proof-of-concept we implemented a web-service that offers the required information to VR applications utilising the proposed framework. In addition we adapted an open-source game with predetermined auditory stimuli to act as a VR application utilising the proposed framework, thus changing the audio sounding based on design-time annotation of the available scenes. The game may be accessed directly through the link: http://karydis.ionio.gr/MIRVRES/

In future work, we plan to examine user satisfaction by contrasting the initial and adapted version of the XNA Arkanoid game. Furthermore, an extension of the proposed framework may incorporate characteristics of multi-user environments that could affect further the selection of musical content, as user-preferences are often contradictory. Finally, provisions may be made for queries concerning not only full tracks but also play lists of tracks and parts of a track such as a theme.

12. REFERENCES

[1] J. Robertson, A. D. Quincey, T. Stapleford, G. Wiggins, "Real-Time Music Generation for a Virtual Environment," presented at the ECAI-98 Workshop on AI/Alife and Entertainment, Brighton, England, 1998.

[2] A. Berndt and K. Hartmann, "The Functions of Music in Interactive Media," in Interactive Storytelling. vol. 5334, U. Spierling and N. Szilas, Eds.: Springer Berlin / Heidelberg, 2008, pp. 126-131.

[3] K. Jørgensen, "On the Functional Aspects of Computer Game Audio," presented at the AudioMostly, Piteå, Sweden, 2006.

[4] M. Naef, O. Staadt, M. Gross, "Spatialized audio rendering for immersive virtual environments," presented at the Proceedings of the ACM symposium on Virtual reality software and technology, Hong Kong, China, 2002.

[5] S. M. Zehnder and S. D. Lipscomb, The Role of Music in Video Games. NJ: Lawrence Erlbaum Associates Publishers, 2006.

[6] K. Collins, Ed., From Pac-Man to Pop Music. UK: Ashgate Publishing, 2008.

[7] K. Collins, Game Sound: An Introduction to the History, Theory, and Practice of Video Game Music and Sound Design: MIT Press, 2008.

[8] K. Jørgensen, "Left in the dark: playing computer games with the sound turned off," in From Pac-Man to Pop Music, K. Collins, Ed., UK: Ashgate Publishing, 2008.

[9] K. Dickinson, Movie Music, The Film Reader: Routledge, 2002.

[10] J. Green, "Understanding the Score: Film Music Communicating to and Influencing the Audience," The Journal of Aesthetic Education, vol. 44, pp. 81-94, 2010.

[11] D. Kaye and J. LeBrecht, Sound and Music for the Theatre: The Art and Technique of Design, 3rd Edition ed.: Elsevier, 2009.

[12] K. Collins, "An Introduction to the Participatory and Non-Linear Aspects of Video Game Audio," in Essays on Sound and Vision, S. Hawkins and J. Richardson, Eds., Helsinki: Helsinki University Press, 2007.

[13] S. Poole, Trigger happy: Videogames and the entertainment revolution. New York: Arcade Publishing, 2000.

[14] P. Lamere, "Social tagging and music information retrieval," New Music Research, vol. 37, pp. 101-114, 2008.

[15] A. Vinciarelli, M. Pantic, H. Bourlar, "Social signal processing: Survey of an emerging domain," Image and Vision Computing, vol. 27, pp. 1743-1759, November 2009 2009.

[16] R. W. Picard, "Affective computing: challenges," Int. J. Hum.-Comput. Stud., vol. 59, pp. 55-64, 2003.

[17] K. Collins, "From Bits to Hits: Video Games Music Changes its Tune," Film International, vol. 13, pp. 4-19, 2004.

[18] J. Futrelle and J. S. Downie, "Interdisciplinary Communities and Research Issues in Music Information Retrieval," in Proceedings of the International Symposium on Music Information Retrieval, Paris, 2002.

[19] M. Slaney, K. Weinberger, W. White, "Learning a metric for music similarity," in Information Society for Music Information Retrieval, 2008, pp. 148-153.

[20] E. Pampalk, "Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns," in nternational Symposium on Music Information Retrieval, 2006.

[21] B. Logan, D. P. W. Ellis, A. Berenzweig, "Toward evaluation techniques for music similarity," presented at the SIGIR 2003: Workshop on the Evaluation of Music Information Retrieval Systems, Toronto, Canada, 2003.

[22] J. J. Aucouturier and F. Pachet, "Music similarity measures: Whats the use?," in International Symposium on Music Information Retrieval, 2003, pp. 157-163.

[23] K. West and P. Lamere, "A model-based approach to constructing music similarity functions," EURASIP Journal on Advances in Signal Processing, pp. 1-10, 2007.

[24] B. McFee, L. Barrington, G. Lanckriet, "Learning similarity from collaborative filters," presented at the International Society for Music Information Retrieval, Utrecht, Netherlands, 2010.

[25] A. Pentland, "Social Signal Processing," IEEE Signal Processing Magazine, vol. 24, pp. 108-111, 2007.

[26] Z. Zeng, M. Pantic, G. Roisman, T. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, pp. 39-58, 2009.

[27] K. Anagnostou. (2009). Arkanoid Part 1. Available: http://videogameslab.wordpress.com/2009/09/14/arkanoid-part-1/

[28] Microsoft. (2004). XNA Programming Environment. Available: https://www.microsoft.com/presspass/press/2004/mar04/03-24xnalaunchpr.mspx

[29] LastFM. (2010). Last.FM API. Available: http://www.last.fm/api

[30] J. B. MacQueen, "Some Methods for Classification and Anlysis of Multivariate Observations", 5th Berkeley Symposioum on Mathematical Statistics and Probability, 1967, pp. 281-297.

[31] F. Gouyon, "A computational approach to rhythm description – Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing". Music Technology Group, Pompeu Fabra University, 2005

[32] M. Davies, M. Plumbley, "Causal tempo tracking of audio", 5th International Symposium on Music Information Retrieval, 2004