Mobile Netw Appl (2011) 16:375–384
DOI 10.1007/s11036-010-0256-4

A System for Mobile Active Music Listening Based on Social Interaction and Embodiment

Giovanna Varni · Maurizio Mancini · Gualtiero Volpe · Antonio Camurri

Published online: 18 July 2010

Abstract Social interaction and embodiment are key issues for future User Centric Media. Social networks and games are more and more characterized by an active, physical participation of the users. The integration in mobile devices of a growing number of sensors to capture users' physical activity (e.g., accelerometers, cameras) and context information (GPS, location) supports novel systems capable to connect audiovisual content processing and communication to users social behavior, including joint movement and physical engagement. In this paper, a system enabling a novel paradigm for social, active experience of sound and music content is presented. An instance of such a system, named Sync'n'Move, allowing two users to explore a multi-channel pre-recorded music piece as the result of their social interaction, and in particular of their synchronization, is introduced. This research has been developed in the framework of the EU-ICT Project SAME (www.sameproject.eu) and has been presented at Agora Festival (IRCAM, Centre Pompidou, Paris, June 2009). In that occasion, Sync'n'Move has been evaluated by both expert and non expert users, and results are briefly presented. Perspectives on the impact of such a novel paradigm and system in future User Centric Media are finally discussed, with a specific focus on social active experience of audiovisual content.

Keywords mobile active music listening · social signal processing · synchronization

G. Varni · M. Mancini · G. Volpe (✉) · A. Camurri
InfoMus Lab-DIST-University of Genova,
Viale Causa 13, 16145 Genova, Italy
e-mail: gualtiero.volpe@unige.it

G. Varni
e-mail: giovannavarni@infomus.dist.unige.it

M. Mancini
e-mail: maurizio.mancini@dist.unige.it

A. Camurri
e-mail: antonio.camurri@unige.it

1 Introduction

The User Centric Media concept "implies that the user will become an active member of the overall media chain by generating, distributing and experiencing high-quality media content" [14]. Key factors for reaching such an active experience and participation of the users include the social and the physical (embodied) dimensions.

The EU-ICT Project SAME (www.sameproject.eu) explores novel paradigms of active, embodied, and social listening to music in context-aware mobile applications [6], i.e., paradigms enabling both single and groups of users to actively mould and reshape sound and music content, based on movement, gesture, and social interaction.

This paper presents a system enabling such new paradigms. The system has been implemented using the SAME Platform for embodied active experience of sound and music content, and in particular the EyesWeb XMI platform for eXtended Multimodal Interaction (www.eyesweb.org).

An instance of the system, named Sync'n'Move, is also presented. Sync'n'Move enables two users to explore a multi-track pre-recorded music piece, as the result of their social interaction. Users interact through the movements they perform by handling a mobile

device. A phase synchronization index is extracted from movement. When users achieve a high level of synchronization, the resulting overall music orchestration and rendering is enhanced, e.g., by increasing the number of music sections, or by enhancing rendering features. Sync'n'Move was presented in live public demonstrations at Agora Festival (IRCAM, Centre Pompidou, Paris, France, June 2009), in a special event dedicated to the EU SAME Project. In that occasion, the system was evaluated by both expert and non expert users.

The remainder of this paper is organized as follows. Section 2 gives an overview of the state of the art in the user-centric media family of systems for active music listening. More details on a scenario where the system finds real application are presented in Section 3. Section 4 presents the system architecture, and its main components: the SAME Platform, the high-level active listening modules, the system's software component for social signal processing, and, in particular, for the analysis of synchronization. Section 5 describes Sync'n'Move, an instance of the system implementing the scenario depicted in Section 3. Results from the evaluation session carried out at the Agora Festival are discussed in Section 6. Perspectives on the impact of such a novel paradigm and system on future User-Centric Media are finally presented in the conclusion, including future directions on social active experience of audiovisual content.

2 Active music listening

Active music listening paradigms enable novel generations of interactive music systems [25], particularly addressing a large public of beginners, naive, and inexperienced users, rather than professional musicians and composers.

Research on active music listening is in its infancy, but initial results begin to be available. Interactive dance-music systems were proposed in [3]: the user full-body rhythmic movements were analyzed in real-time and compared with the beat of a song (extracted from the MIDI music signal). The result was the interactive control of the quality of the orchestration, of the melodic pitch (de)tuning, and of the synchronization of the different voices of the song. In short, the better the movement (and in particular the more synchronized the movement with the song beat), the better the rendering of the music heard. A set of applications on "dance karaoke" were developed and presented in a number of public events and prototypes. Goto [11] proposed a content-centric system for intervening on pre-recorded music with signal processing techniques to select sections, skip and navigate parts of the recording. He developed some original real-time signal processing techniques on the audio signal, but his approach lacks user-centric aspects: standard GUIs on a PC are used to apply signal processing to audio files. Pachet [19, 20] at Sony CSL proposed a similar approach to active music listening based on the audio mixing paradigm. Another system proposed by Pachet, the Continuator, is an auto-reflexive system capable to build music improvisation experiences in which the system plays a role of a companion of the user in music sessions. A few aspects of content- and user-centric components are implicitly introduced in the parameters settings of the Continuator. A music piano keyboard is the user interface, so that this is a system for music performance practice, rather than for active music listening.

The active music listening system presented in this paper is grounded on the analysis of social interaction, and in particular on inter-personal synchronization. This opens novel directions towards the automated measurement and exploitation in user-centric media applications of emotional engagement and empathy, based on high-level synchronization of expressive features. Research works addressing synchronization in human-human or human-machine interaction are available: for example, some approaches exploit synchronization to allow the users to compete in dancing or to learn to play virtual instruments; others aim at developing machines able to assist motor impaired people in rehabilitative sessions. Leman et al. [16] reworked the concept of social music game: the movement beat of multiple users is extracted and compared with the beat of the music the users are listening to. Users can compete among them or collaborate to win the game. Kirschner and Tommasello [13] argue that children tend to learn to follow tempo in a collaborative social context; in their work they present a methodology for measuring, by means of synchronization, the bias between the reference and the children beats timing. In Robotics several assisting synchronization-based systems were developed: an example is the Walk-Mate robot, a virtual locomotion collaborative walking system able to support the walking of Parkison's disease and hemiplegia patients. It instructs user gait motion providing acoustic stimuli and evaluates the user improvements by a phase difference between the stimuli and the gait timings [18].

Interfaces promoting collaboration were conceived to drive audio content generation or manipulation in multi-user scenarios. Stokholm and Pasquier [26], for example, implemented a system which mixes audio representations of the moods of several users in a

room to increase collaboration and empathy among users. The Audio Explorer system enables users to concurrently modify the audio mixing of a piece of music downloaded from the Web and to share the resulting content [7]. Another example is Herd It [1], a multi-user game running on the Facebook social network, that collects tags about music and in which scoring depends on consensus between a large group of listeners.

Whereas most of the current systems address synchronization and social interaction at a low level (e.g., based on beat extraction), the availability of automated and real-time analysis techniques of the high-level expressive and the emotional processes in the users, including their entrainment, empathy, and other social signals [21] can be very useful to understand the users' behavior, and to use this information to shape joint music listening experiences [4, 8, 12]. Recent research started investigation in this direction [27], that will be further carried out in the EU-ICT FET Project SIEMPRE.

3 Scenario: social active music listening

Consider a non-professional music band composed by young people. They frequently meet to play together and they have contacts with many friends both where they live and internationally. For example, they have an account on a social network (e.g., MySpace) where they periodically upload and share with their friends the music pieces they played and recorded. They tag their recordings with consolidated metadata concerning, for example, author, duration, and music genre. They can also put textual descriptions of their pieces and can receive comments from their friends. The system presented here allows them to add a further, fundamental dimension, that is, *embodiment*: "the human body is a mediator between meaning formation activities at the mental level, and musical signals at the physical level", as stated by De Bruyn et al. in [9]. The members of the band become, the members of the band can tag their pieces with metadata concerning embodiment (e.g., the danceability of the piece, cross-modal music/gesture representations, such as whether the piece is 'impulsive' or 'quick' or 'hesitant') and expressive, emotional content (e.g., whether the piece is 'introvert', 'solemn', 'joyful', or whether it expresses anger).

Besides just downloading and listening to the music, the friends can access to a social, active experience of the piece, through their mobile phones running the clients of the SAME end-to-end platform. For example, they can meet (at home, in the street, in a disco or even virtually on a social network) and run Sync'n'Move from their mobiles (Sync'n'Move is described in more details in the following sections). They select a piece based on the embodiment and expressive metadata (or the SAME server can recommend a music piece based on their user's profiles and context information). They can then freely move their mobile devices trying to synchronize their gestures both at the physical level (e.g., gestures have similar profiles of kinematical features) and at the expressive, emotional level (e.g., gestures are characterized by similar profiles of expressive features, such as, for example, energy, impulsiveness, fluidity). The more they synchronize the more the music output is enhanced by adding sound tracks and rendering features, e.g., from a poor monophonic audio to a full polyphonic experience. They can even synchronize and converge on different expressive performances of the same piece (e.g., an 'introvert' or a 'solemn' one).

Audio is streamed directly from the SAME server to the mobile devices. However, if the friends meet in a place where more sophisticated audio rendering systems are available (e.g., a HiFi at home, a professional audio system at a discotheque or public space, a full featured Wave Field Synthesis system for 3D audio experience), the audio stream can be redirected to such a more sophisticated system in order to dynamically achieve the best conditions for the active experience.

The friends can leave comments about their experience both in terms of textual messages and as annotations concerning their active experience (e.g., their feelings) expressed through the metadata for embodiment and expressive content. Comments are available to other friends or other people who can also try and comment a similar experience.

4 System architecture

The system architecture (see Fig. 1) consists of modules running on client mobile devices and modules running on servers. The modules on the clients perform input/output operations (input of multimodal data and output of multimedia content) and computationally cheap processing. The modules on the servers perform computationally expensive processing, including feature extraction and social interaction analysis, controlling in real-time the processing of multimedia streams, and search and selection of multimedia content based on metadata (including expressive and context-related metadata).

On the client side, the *data acquisition module* captures: (i) data from the on-board sensors (e.g., ac-

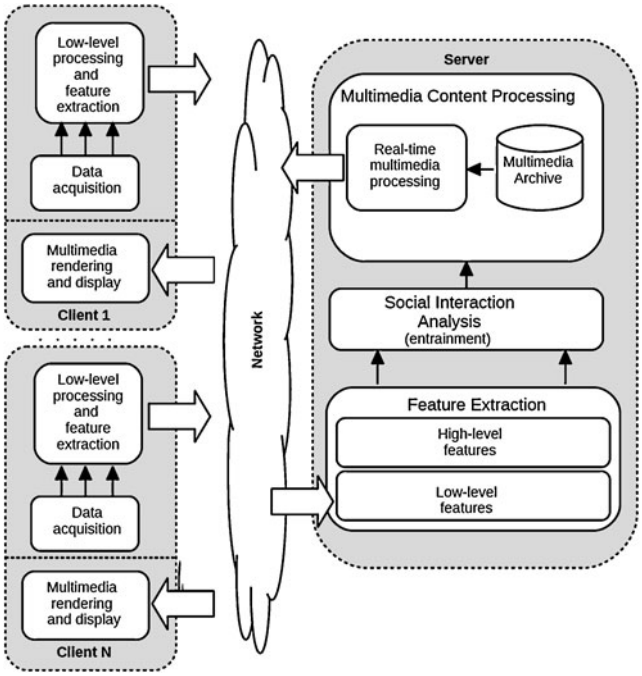

Fig. 1 The system architecture

celerometers, camera, microphone, GPS, temperature) and (ii) text and other information provided by the user through the keypad and the buttons the mobile device is endowed with. The client *module for low-level processing and feature extraction* performs on such data operations such as filtering and extraction of simple features, e.g., the average value of acceleration in a time window. Such a low-level information, either raw signals or pre-processed (filtered) data or time series of low-level features is sent to one or more servers through the network. The client *module for multimedia rendering and display* receives multimedia streams from a server and renders them on the mobile device.

On the server side, the *module for feature extraction* is based on the layered conceptual framework developed by Camurri and colleagues [4] to analyze and model expressive gesture. This framework—originally conceived for analysis of full-body movement—has been adapted and refined in order to be applied on data referring to the movement of handheld devices. The extracted features range from low-level physical measures (e.g., acceleration, velocity, and their statistics), towards overall higher-level gesture information including gesture features (e.g., motion energy, fluidity, impulsiveness, directness), information describing semantic properties of gestures (e.g., greeting, pointing, grasping objects), the activities the user is performing (e.g., walking, running, driving), and the expressive qualities of her movement (e.g., emotion, affective attitudes). These are obtained by applying machine learning techniques to the overall gesture features. Audio features (e.g., describing the prosody of the voice of the user or trying to characterize possible environmental sounds) are obtained by the audio signals coming from the microphone of the mobile device. Contextual features (e.g., the user is at home, office) are recovered from the approximated GPS location, the activity the user is performing and audio features (e.g., environmental sounds). Finally, social features are either analyzed directly (for example, synchronization) or are inferred by the expressive and contextual features: e.g., the user is in a happy emotional state while interacting with other people at a friend's home during a party.

The *module for multimedia content processing* is fed with the high-level expressive and social features and manipulates and molds the selected multimedia content before streaming it to the client for rendering. This module includes on the server a multimedia archive, associated with a module for selection of multimedia content. In this way, the information describing the context and the expressive and social behavior of the user is matched with the pre-annotated metadata stored in the multimedia archive, and recommendations are provided to the user in order to select an appropriate multimedia content for the active experience.

In the following, the *social interaction analysis module* is described in more details, with particular reference to the algorithms for measuring synchronization.

4.1 The module for analysis of social interaction

The main goal of the module for analysis of social interaction is to measure synchronization between users. This Section presents the algorithms developed to measure in real-time synchronization, intended as the synchronization of the movement behavior between two users moving two mobile phones in a physical space. The adopted approach is based on analysis of Phase Synchronization (PS) in complex systems [22]. Each user is described by an n-dimensional state vector in which the n dimensions are n features like, for example, trajectories of body segments, amount of motion, audio descriptors, and so on. On this assumption, interaction is addressed considering how the state vectors evolve together in time and extracting, from this joint dynamics, indices, such as, e.g., a Phase Synchronization index of the *global* behavior of the system. For example, in the simplest case (adopted in Sync'n'Move), the two users are simply described by the 1-dimensional state vectors containing the absolute value of the 3D

acceleration measured by on-board accelerometers the users' mobile phones are endowed with.

Phase Synchronization is measured using mathematical techniques based on the recurrence property of dynamical systems [23]: more specifically, Recurrence Plots (RPs) [10] and Recurrence Quantification Analysis (RQA) [17, 28]. These techniques provide qualitative and quantitative information on the systems' dynamics and their interrelations in terms of trajectories in a chosen (physical or reconstructed) features space. An RP is a time-time binary colorimetric plot displaying all the time instants in which recurrences in the state of a system are observed, whereas RQA quantifies the graphical patterns occuring in an RP.

Let us consider two users, from here on identified by their state vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. The RP for the first user (the same formula is also valid for the second one by replacing $\mathbf{x}$ with $\mathbf{y}$) is defined through the recurrence matrix:

$$R_{i,j}(\varepsilon) = \Theta\Big(\varepsilon - \| \mathbf{x}_i - \mathbf{x}_j \|\Big) \qquad i, j = 1...N \tag{1}$$

where, $\mathbf{x}_{i,j} \in \mathbb{R}^n$ are the user states at times i and j, N is the number of the states (number of samples), ε is a closeness threshold, $\|\|$ and Θ are a norm (e.g., the Euclidean norm, the minimum or maximum norms can be adopted) and the Heaviside function, respectively. $R_{i,j}$ are binary values (0 or 1) according to whether $\mathbf{x}_i$ is *close* or not to $\mathbf{x}_j$.

By applying RQA to an RP, the probability $\hat{p}(\varepsilon, \tau)$ of recurrence at a certain state after some time τ is computed [24]. The estimate of this probability can be written as:

$$\begin{aligned}\hat{p}(\varepsilon, \tau) &= \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} R_{i,i+\tau}(\varepsilon) \\ &= \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} \Theta(\varepsilon - \| \mathbf{x}_i - \mathbf{x}_{i+\tau} \|)\end{aligned} \tag{2}$$

Again, obviously, the same formula is valid for the second user by replacing $\mathbf{x}$ with $\mathbf{y}$. Finally, the PS between the two users is computed with the Correlation Probability of Recurrence (CPR), defined as:

$$CPR = \langle \bar{p}_{\mathbf{x}}(\varepsilon, \tau) \bar{p}_{\mathbf{y}}(\varepsilon, \tau) \rangle \tag{3}$$

where $\bar{p}_{\mathbf{xy}}(\varepsilon, \tau)$ are the functions $\hat{p}(\varepsilon, \tau)$ normalized to zero mean and unitary standard deviation.

The social interaction analysis module can be extended for analyzing further social signals, such as for example analysis of functional roles (e.g., leadership).

5 An instance of the system architecture: Sync'n'Move

Figure 2 sketches out the architecture of Sync'n'Move. Two users freely move their mobile phones and their hand or body (the mobile may be kept either in a hand or weared in a pocket) acceleration is detected and measured by the tri-axial accelerometer onboard the mobile phones. The CPR index of Phase Synchronization is extracted from their gesture: every time the users succeed in synchronizing (the corresponding index is high), the music orchestration and rendering is enhanced. If instead the users do not synchronize (i.e., the Phase Synchronization index is low), the music gradually degradates, by deleting progressively music sections. Further possibilities can be added, e.g., by reducing timbral quality of sections and sound spatialization rendering.

With respect to the overall system architecture in Fig. 1, the Sync'n'Move instance implements the following modules: on the client side, it extracts the raw acceleration from the on-board accelerometers (*acquisition module*) and renders the music output (*multimedia rendering and display module*). Pre-processing and feature extraction is performed on the server side. The server applies pre-processing and filtering techniques (*low-level features*) and performs analysis of synchronization (*social interaction analysis module*). It also performs audio processing (*real-time multimedia processing module*). Figure 2 shows the two main modules composing Sync'n'Move on the server side: *the data acquisition module* and *the feature extraction & audio processing module*. The former implements pre-processing and filtering techniques (i.e., the *low-level feature extraction module*); the latter implements the analysis of social interaction and the audio processing chain (i.e., the *social interaction analysis module* and the *real-time multimedia processing module*).

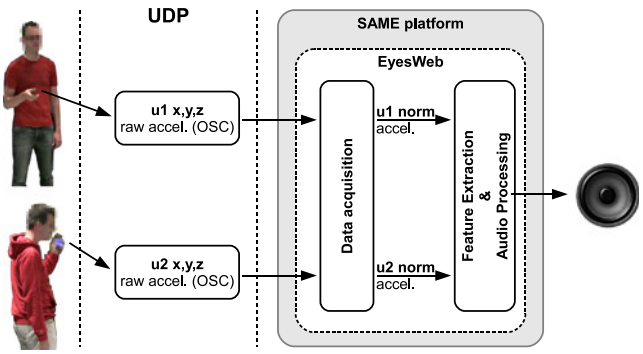

Fig. 2 Architecture of Sync'n'Move

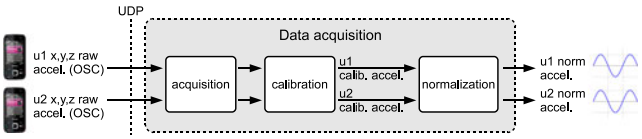

Fig. 3 The data acquisition module

The data acquisition module This module (see Fig. 3) acquires, calibrates, and computes the normalized acceleration captured by the mobile phones the users are moving.

Each mobile runs a Python script that collects data from the accelerometer and creates an OSC packet in the form:

```
/synchronizer ax ay az
```

The above packet is sent via UDP to the SAME platform running EyesWeb where the raw accelerations on the 3 axes are extracted from the OSC packet (*acquisition* block in Fig. 3). The *calibration* and *normalization* blocks are necessary since every accelerometer has a different ground reference, that is, the *max* and *min* values of **g** change on every sensing axis. The operations performed by the *calibration* block on the x axis are here reported. The same computation is necessary also on the other two axes.

$$A_{C_x} = A_{raw_x} - IMD_x; \quad IMD_x = \frac{gx^+ + gx^-}{2}; \tag{4}$$

where:

- IMD_x is the half of the *Inter Maxima Difference*, that is, the half of the difference between the maximum **g** measured on x^+ and x^-;
- A_{C_x} is the calibrated acceleration on the x axis, that is, the acceleration obtained by subtracting the IMD_x from the raw acceleration on x, that is, A_{raw_x};

Finally, the *normalization* block computes the absolute value of acceleration and nomalizes it in the range [0, 1] subtracting **g** in order to ignore it:

$$A_{\text{normalized}} = \frac{\sqrt{A_{C_x}^2 + A_{C_y}^2 + A_{C_z}^2} - \mathbf{g}}{A_{MAX}}; \tag{5}$$

where A_{MAX} is the maximal absolute value of acceleration detected by the phone.

The feature extraction & audio processing module Figure 4 shows the details of the feature extraction & audio processing module. This module is responsible for computing the Phase Synchronization index, which is used for controlling audio processing. Starting from

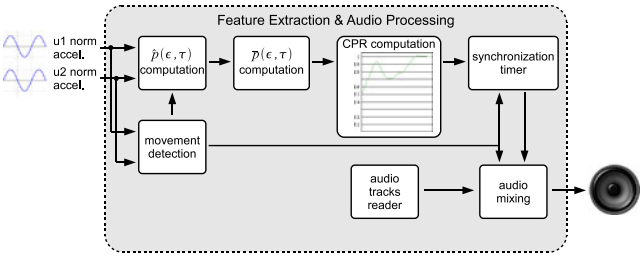

Fig. 4 The feature extraction & audio processing module

the normalized accelerations ($A_{\text{normalized}}$), the probabilities of recurrence $\hat{p}_x(\varepsilon, \tau)$ and $\hat{p}_y(\varepsilon, \tau)$ are computed. These are then normalized to obtain the probabilities $\bar{p}_x(\varepsilon, \tau)$ and $\bar{p}_y(\varepsilon, \tau)$ having zero mean and unitary standard deviation. The next step is the computation of CPR. The final output is an audio content produced by the *audio processing* block. This content changes according to the synchronization degree between the users. The following three cases can occur:

- no audio: the users are not interacting at all, that is they are not moving their mobile phones. In this case the *movement detection* block detects that the two accelerations are equal to zero and inhibits audio generation;
- metronomic audio: (i) only one user is moving or (ii) both are moving but they are not synchronized. In the first condition, the *movement detection* block detects that just one of the accelerations is different from zero and enables the generation of a metronomic section in the audio output, e.g., the charleston instrument. In the second condition, the CPR is computed but it is too low to allow the generation of the full audio output.
- full audio: the two users are moving in a synchronized way. The CPR assumes a high value (almost one) and the *synchronization timer* measures the time along which the two users keep synchronized. According to the duration of this time, new sections are added to the audio content: the longer is the synchronization time the larger is the number of the enabled instruments, e.g., drums, bass and guitar, voice. No changes are applied to, for example, the music rhythm, intonation, volume and so on.

An example of how Sync'n'Move works can be found at: www.sameproject.eu/Demos.

6 Technology infrastructure

The system is implemented on the SAME Platform, an end-to-end framework for distributed dynamic

Fig. 5 *Upper row:* users trying Sync'n'Move at Agora Festival, IRCAM, Paris, France, June 2009. *Lower row:* snapshots of the real-time processing performed by EyesWeb XMI. *Left panel* shows the probabilities of recurrence of the accelerations, *right panel* shows RPs

processing and streaming of synchronized multimodal channels. The SAME platform architecture consists of multiple clients connected to a variable number of servers running software environments, such as for example EyesWeb XMI [5] and MAX MSP (http://www.cycling74.com).

Clients run on mobile devices. They capture and communicate to the servers in real-time the data obtained from the on-board sensors (e.g., accelerometers, camera, microphones), show graphic interfaces, provide visual feedback, and play audio content streamed from the servers.

Servers manage audio collections (annotated with metadata), provide content recommendation based on user's profiles and context-aware information, and perform computations needing computational resources that are not available (yet) on the client devices, e.g., audio modulation or gesture analysis. In future mobiles it is expected a migration of computation from servers to clients.

Servers receive from clients low-level data about the gestures users performed and context information. At the end of their computation, data is sent back to clients: e.g., a piece of music is selected from a library depending on the user's mood and then sent back to the user's mobile device; a new version of a pre-recorded MP3 is computed and streamed to a group of users that collaboratively contributed to remix it in real-time.

7 System validation

Sync'n'Move was tested during a public event organized by the EU SAME Project at IRCAM (Centre Pompidou, Paris) as part of the multidisciplinary Agora Festival in June 2009. Figure 5 shows some users trying the system.

An evaluation was carried out via an anonymous assessment questionnaire conceived by the SAME Consortium and collecting information from the event attendees on nine prototype systems designed and developed by the project partners. The questionnaire

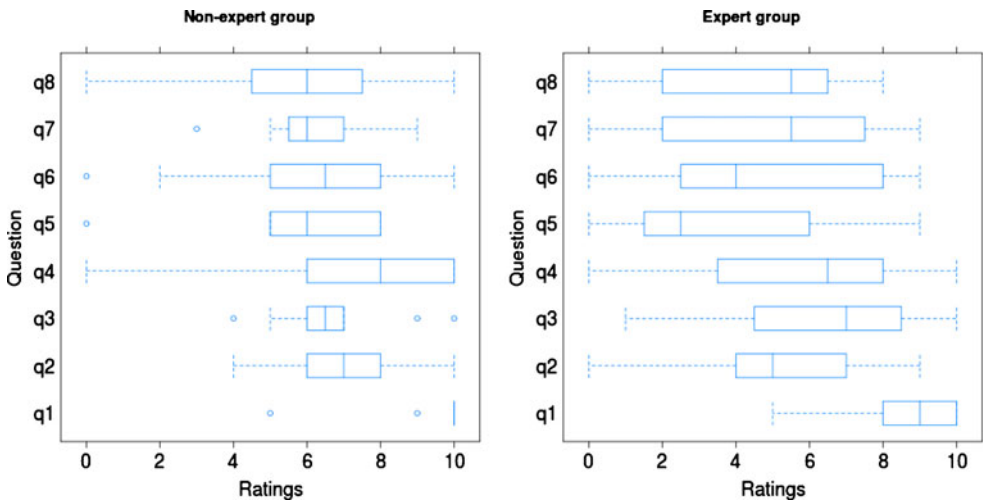

Fig. 6 Box plots of the ratings of the non-expert users (*left panel*) and of the expert users (*right panel*). The questions on the y-axis stand for: *q1*-understanding, *q2*-control, *q3*-interaction, *q4*-fun, *q5*-interest, *q6*-future, *q7*-engagement, *q8*-enjoyment. The median of each distribution is marked by the *horizontal line* within the boxes. The *upper and the lower hinges* are depicted by the edges of the boxes. The *inner fences* are depicted by the upper and lower whiskers coming out from each box. Possible outliers are marked with *empty dots*

both gathered general data about participants, such as age, gender, nationality, occupation, musical skills and addressed questions about the prototypes. Specific questions were designed for each prototype. In particular, questions on Sync'n'Move concerned the following items: *understanding*, *control*, *interaction*, *fun*, *interest*, *future*, *exploitation*, *engagement*, and *enjoyment*. Participants were asked to express their ratings on 11-steps scales ranging from *not at all* to *very*. A blank space was left in the last page to collect comments and suggestions. Partially filled up questionnaires were also taken into account.

Twenty-two (22) people (19 male and 3 female coming from european and extra-european countries) evaluated Sync'n'Move. Mean age of people was 33.7y (from 18y to 64y). All the participants were volunteered for this study and they were only asked to have a spontaneous behavior as much as possible. Before their performance, they were provided with a short demonstration of how the system works. The participants were classified in two groups: experts and non experts. The experts group was composed by 12 subjects having attested expertize in the musical field such as composers and professional musicians. Box plots were computed for each group (see Fig. 6). The x-axis shows the ratings from 0 to 10, whereas the y-axis shows the questions concerning the items. From visual inspection of the plots, it is clear that Sync'n'Move was very easy to understand for both groups of users. Globally, the ratings provided by expert users are more largely spread than those of the non expert users. A Mann-Whitney test with a post hoc Bonferroni correction was used to verify whether the two groups can be considered coming from the same population with respect to the items. The post hoc Bonferroni correction of the level of statistical significance was needed to address the problem of multiple comparisons. The corrected level of statistical significance was $p < 0.00625$. The results showed that there is not evidence that the users were coming from different populations. However, effect size calculation revealed that medium effect size differences occur for the items *interest* ($r = 0.37$) and *pleasure* ($r = 0.30$). These differences can be ascribed to a greater perception by expert users that the audio tempo genereated by Sync'n'Move was not matching in time with the tempo chosen by the users. Moreover, Sync'n'Move exploits very simple music content and interaction mechanisms, whereas expert users probably expect more complex paradigms and content. Finally, evaluation of the degree of involvement with Sync'n'Move was performed. For this sake, reference was done to the layered Immersion Model of Brown and Cairns [2]. This model was conceived to investigate the intensity of immersion in video games, however it can be used to describe user experience in the more general HCI context. Brown and Cairns detail three levels of immersion, *engagement*, *engrossment*, and *total immersion*, that can be reached by players overcoming barriers. The analysis carried out for Sync'n'Move deals only with the first level of immersion: engagement. Barriers for reaching engagement are *access* (i.e., the users preferences and liking), *control*, and *feedback*. Spearman correlation coefficients were used to describe the relation between the pairs of items *engagement-control*, *engagement-enjoyment*, and *engagement-fun*. *Enjoyment* and *fun* are labels for *access*. The significant correlations between *engagement-enjoyment*, $\rho = 0.69$, $p = 0.05$ and *engagement-fun*, $\rho = 0.76$, $p = 0.05$ for the non expert group indicate and confirm a relationship between these items and support the conclusions drawn above.

8 Conclusion

This paper presented a system enabling new paradigms of social active music listening, with a special focus on embodiment. Social interaction and embodiment emerged as key factors for future applications envisaging social experience of multimedia content. An instance of the system, Sync'n'Move, has been introduced and discussed, with particular reference to such a key factors. Further qualitative and quantitative user experience evaluation of Sync'n'Move, involving a larger quantity of test users, is planned at the final event of the SAME project in November 2010.

Social interaction and embodiment are expected to play a significant role on future research on User Centric Media and the Future Internet. On the one hand, the worldwide spreading of social networks and Internet games bears witness of the importance of the social dimension. Nevertheless, many existing multimedia interactive systems and Internet applications are still intended for a single user and social interaction is often neglected. Social networks are mainly based on sharing of static textual and audiovisual content, whereas realtime interaction between users, immersiveness, and sense of presence are far to be fully reached.

On the other hand, active experience requires embodiment. As the current trend of game industry shows, movement and gesture—either performed using dedicated hand-held devices or captured by motion capture and tracking systems—are extremely important for the development of games that actively engage their users. Consider, for example, Nintendo Wii, Sony

PlayStation, and the recently announced Microsoft Natal Project for the future Xbox.

In the Future Internet users will play a crucial role both as individuals (endowed of a body) and in the social dimension [15]. In a User Centric perspective, one of the key factor and major research challenges for the Future Internet will be the convergence of technologies enabling the development of social embodied networks, supporting active and shared experience of multimedia content.

Such a convergence will open novel perspectives in many different application fields, e.g., future networked media, performing arts applications, therapy and rehabilitation, novel paradigms and systems for active experience of cultural heritage. In this framework, promising research directions include the extension of the system and the development of futher prototypes going beyond Sync'n'Move, e.g., in the direction of (i) improving the control that users have on generation and manipulation of multimedia content, (ii) extending the concept of active experience towards full multimedia content, (iii) investigating algorithms for analysis of synchronization of multiple users, (iv) developing computational models and algorithms towards complex synchronization phenomena such as empathy.

Acknowledgements The authors would like to thank Paolo Coletta for EyesWeb support and development and Alberto Massari for creating Python scripts for reading and transmitting accelerometer data and audio streams for Nokia phones. The authors also thank Norbert Marwan for precious suggestions on recurrence and Carlo Chiorri for precious discussion on statistics.

The research described in this paper is partially supported by the EU FP7 ICT SAME project, and also benefits from the preliminary results obtained in the EU FP7 ICT I-SEARCH project, specifically in advances on the retrieval of sound materials basing on user interaction; future work within I-SEARCH will include the extension of the paradigm to audiovisual search and retrieval.

References

1. Barrington L, O'Malley D, Turnbull D, Lanckriet G (2009) User-centered design of a social game to tag music. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, Paris, ACM New York, NY, USA
2. Brown E, Cairns P (2004) A grounded investigation of game immersion. In: Proceedings of Conference on Human Factors in Computing Systems (CHI 2004), pp 1297–1300
3. Camurri A (1995) Interactive dance/music systems. In: Proceedings International Computer Music Conference (ICMC-95), The Banff Centre for the arts, Canada, ICMA-Intl.Comp.Mus.Association, pp 245–252
4. Camurri A, De Poli G, Leman M, Volpe G (2005) Toward communicating expressiveness and affect in multimodal interactive systems for performing art and cultural applications. IEEE Multimedia Magazine 12(1):43–53
5. Camurri A, Coletta P, Varni G, Ghisio S (2007) Developing multimodal interactive systems with EyesWeb XMI. In: Proceedings of the 7th International Conference on New Interfaces for Musical Expression, Genova, ACM New York, NY, USA, pp 305–308
6. Camurri A, Volpe G (2008) Active and personalized experience of sound and music content. In: Proceedings of 12th IEEE International Symposium on Consumer Electronics, IEEE Press
7. Camurri A, Volpe G, Vinet H, Bresin R, Fabiani M, Dubus G, Maestre E, Llop J, Kleimola J, Oksanen S, Välimäki V, Seppanen J (2009) User-centric context-aware mobile applications for embodied music listening. In: Proceedings of the 1st International ICST Conference on User Centric Media. ISBN 978-963-9799-84-4, Venice, Italy
8. Clayton M, Sager R, Will U (2004) In time with the music: the concept of entrainment and its significance for ethnomusicology. ESEM counterpoint 1:1–82
9. De Bruyn L, Leman M, Moelants D (2008) Quantifying children's embodiment of musical rhythm in individual and group settings. In: Proceedings of the 10th International Conference on Music Perception and Cognition, Sapporo, Japan
10. Eckmann JP, Kamphorst SO, Ruelle D (1987) Recurrence plots of dynamical system. Europhys Lett 5:973–977
11. Goto M (2007) Active music listening interfaces based on signal processing. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), IV-1441-1444
12. Keller PE (2008) Joint action in music performance. In: Morganti F, Carassa A, Riva G (eds) Enacting Intersubjectivity: a cognitive and social perspective on the study of interaction. IOS Press, Amsterdam, pp 205–221
13. Kirschner S, Tomasello M (2009) Joint drumming: social context facilitates synchronization in preschool children. J Exp Child Psych 102:299–314
14. Laso-Ballesteros I, Daras P (eds) (2008) User centric future media internet. EU Commission
15. Laso-Ballesteros I, Daras P (eds) (2009) User centric media in the future internet. EU Commission
16. Leman M, Demey M, Lesaffre M, van Noorden L, Moelants D (2009) Concepts, technology and assessment of the social music game 'Sync-in Team'. In: Proceedings of the 12th IEEE International Conference on Computational Science and Engineering Vancouver, BC, Canada. IEEE Computer Society
17. Marwan N, Romano MC, Thiel M, Kurths J (2007) Recurrence plots for the analysis of complex systems. Physics Reports 438:237–329
18. Miyake Y (2009) Interpersonal synchronization of body motion and the walk-matewalking support robot. IEEE Trans Robot Autom 25(3):638–644
19. Pachet F, Delerue O (2000) On-the-fly multi track mixing. In: Proceedings of 109th AES Convention, Los Angeles, USA
20. Pachet F (2004) Creativity studies and musical interaction. In: Delige I, Wiggins G (eds) Musical Creativity: current research in theory and practice. Psychology Press
21. Pentland A (2007) Social signal processing. IEEE Signal Process Mag 24(4):108–111
22. Pikovsky A, Rosenblum MG, Kurths J (2001) Synchronisation: a universal concept in nonlinear sciences. Cambridge University Press, Cambridge

23. Poincaré H (1890) Sur le probléme des trois corps et les équations de la dynamique. Acta Mathematica 13:1–271
24. Romano MC, Thiel M, Kurths J, Kiss IZ, Hudson JL (2005) Detection of synchronisation for non-phase coherent and non-stationarity data. Europhys Lett 71(3):466–472
25. Rowe R (1993) Interactive music systems: machine listening and composition. MIT Press, Cambridge MA
26. Stockholm J, Pasquier P (2009) Reinforcement learning of listener response for mood classification of audio. In: Proceedings of the 12th IEEE International Conference on Computational Science and Engineering Vancouver, BC, Canada. IEEE Computer Society
27. Varni G, Camurri A, Coletta P, Volpe G (2009) Toward real-time automated measure of empathy and dominance. In: Proceedings of the 12th IEEE International Conference on Computational Science and Engineering Vancouver, BC, Canada. IEEE Computer Society
28. Zbilut J, Webber CL Jr (1992) Embeddings and delays as derived from quantification of recurrence plots. Phys Lett A 5:199–203