EAI Endorsed Transactions
on Creative Technologies
Research Article ICST.ORG

Social retrieval of music content in multi-user performance

Maurizio Mancini*,1, Gualtiero Volpe[1], Giovanna Varni[1], Antonio Camurri[1]
[1]Casa Paganini - InfoMus, DIBRIS, University of Genoa, Italy

Abstract

An emerging trend in interactive music performance consists of the audience directly participating in the performance by means of mobile devices. This is a step forward with respect to concepts like active listening and collaborative music making: non-expert members of an audience are enabled to directly participate in a creative activity such as the performance. This requires the availability of technologies for capturing and analysing in real-time the natural behaviour of the users/performers, with particular reference to non-verbal expressive and social behaviour. This paper presents a prototype of a non-verbal expressive and social search engine and active listening system, enabling two teams of non-expert users to act as performers. The performance consists of real-time sonic manipulation and mixing of music pieces selected according to features characterising performers' movements captured by mobile devices. The system is described with specific reference to the SIEMPRE Podium Performance, a non-verbal socio-mobile music performance presented at the Art & ICT Exhibition that took place in Vilnius (LI) in November 2013.

Keywords: personalised social media experience in mobile devices, embodied cooperation, expressive and social features, music retrieval
Received on 19 June 2014, accepted on 16 September 2014, published on 2 June 2015

doi: 10.4108/ct.2.3.e1

1. Introduction

Along with the widespread dissemination of smartphones or, in general, of mobile devices endowed with increasing computational power, users are enabled to participate more actively in creative experiences by means of their natural (expressive and social) behaviour measured from data mobile devices capture. In interactive music performance a peculiar expression of such an emerging trend consists of the audience directly participating in the performance by means of mobile devices. This is obtained, for example, by distributing specific apps during the performance (e.g., see [1] for a technique for optimised distributions of apps in an audience) or by exploiting existing apps or social media (e.g., see [2] for an example of real-time sonification of tweets).

This is a step forward with respect to concepts like active listening and collaborative music making. In active listening, users (e.g., performers and/or customers) are enabled to adapt/modify the music content they are listening to. In collaborative music making, a group of users is enabled to create and play together a (new) music piece. Starting from such concepts and somewhat combining them, non-expert persons (e.g., members of an audience) are enabled to directly participate in a creative activity such as performing music: they can either become performers or collaborate with professional performers in a creative joint action.

Developing such novel forms of music performances requires the availability of technologies for capturing and analysing in real-time the natural behaviour of the users/performers, with particular reference to non-verbal expressive and social behaviour. Indeed, music is a media for expressing emotional states, moods, personality traits and stances [3], and music performance is an inherently expressive and social activity.

This paper presents a prototype of a non-verbal expressive and social search engine and active listening system, enabling two teams of non-expert users to act as performers, where the performance consists of real-time sonic manipulation and mixing of music pieces selected according to features characterising performers' movements captured by mobile devices.

The system is described with specific reference to how it was used for the *SIEMPRE Podium Performance*, a non-verbal socio-mobile music performance Casa Paganini - InfoMus presented at the Art & ICT Exhibition[1] that took place in Vilnius (LI) in November 2013. The exhibition was included in the program of the EU ICT 2013 event, organised by the European

[1] http://ec.europa.eu/digital-agenda/en/ict-2013-exhibition

*Corresponding author. Email:maurizio.mancini@unige.it

Commission in partnership with the Lithuanian Presidency of the Council of the EU, the Europe's biggest digital technology event where the most recent research and developments on information and communication technologies are presented. The performance is a joint action involving two teams of three performers using smartphones. Performers are not expert musicians and may represent members of a generic audience. They have to compete by shaking rhythmically their smartphones or by moving/dancing while wearing tightly their smartphones to their body, to choose the music to be played actively in the common environment. Performers' movement information is acquired by the gyro and 3-axes accelerometers embedded in the smartphones. The team who is expressing the most coherent expressivity in terms of movement tempo, fluentness and motion energy, is the one that makes music emerge. That is, we aim to model the following analogy: the more a group of music players moves in a coherent way during performance, the more enjoyable (for the both audience and players) a music performance is.

After discussing some related work in Section 2, the concept of the performance is introduced in Section 3. Section 4 presents the architecture of the prototype followed in Sections 5 and 6 by the expressive and social feature extraction components as they were used in the SIEMPRE Podium Performance. Finally, Section 7 presents the music database for the performance and the components for song selection and interaction management.

2. Related work

This paper focuses on how the non-verbal expressive joint action of a group of people extends the action affordances that are possible for a single individual [4]. The trend of game industry (e.g., Nintendo and Microsoft) shows that even a simplified non-verbal expressive joint action, i.e., one that does not take into account measures of emotional and social content in a constrained ecological setting, can be successful in engaging users in video games (e.g., see [5][6][7]). However, merging non-verbal expressive joint action and mobile technologies is still under-investigated and opens new perspectives for future *User Centric Media* [8].

Systems that explicitly measure social signals and perform analysis of social interaction are more and more emerging in the HCI community (e.g., see [9][10][11] and [12] for a survey on social signal processing), but are still missing embodied cooperation and collaborative features, e.g., they do not allow multiple users to collaborate in performing common tasks or reach common goals. Conversely, existing music making and information retrieval systems do not yet take into account features from social signal analysis, e.g., entrainment, empathy, and so on.

2.1. Socio-mobile signal analysis

Several works used mobile devices like smartphones and tablets to analyse social relationships between users. Lukowicz et al. [13] demonstrated that social information they define as "behavior demographics", that is, preferences and risks of segments of a population, could be computed by looking at people's mobile phone traces. In another work (see [14]) they investigated the evolution of social relationships in a student dormitory by monitoring the position of 70 users every 6 minutes over nine months. They collected data from mobile phones such as phone calls and text messages between people living in the dormitory, physical proximity (detected by Bluetooth signals) and position in the dormitory (detected by WiFi signals). This data was used to estimate friendships and their evolution over time (e.g., persons are more likely to become friends if they frequently share the same space) and to monitor social activities. In [15], different types of conversations between users, mainly brainstorming and decision making, were distinguished using mobile wearable devices embedding a microphone and a radio transmitter that communicated the recorded data to radio hotspots. When concerning a better understanding of interpersonal creative interaction (such as playing music), authors in [16] state that detecting finer-grained physiological process is determinant. Their MobileMuse, an unobtrusive sensor package for mobile physiological signal acquisition, aims at collecting seamlessly temperature, GSR, pulse oximetry, and triaxial accelerometer to characterise critical component of social interaction such as emotional contagion.

2.2. Collaborative music systems

Collaborative music systems have a twofold focus: music making and music listening. Traditionally conceived for laptops and workstations, mobile versions of these systems recently appeared.

Concerning music making, Blaine and Fels [17] claim that such systems commonly restrict musical control, facilitating novices' participation in the musical experience. Further, as Jordà argues [18], multi-user instruments facilitate responsiveness and interaction between each performer and the instrument, and also between performers. Among music making systems, the ReacTable [19] allows a group of people to share control of a modular synthesiser by manipulating physical objects on a round table; Audiopad is a composition and performance instrument for collaborative electronic music which tracks the positions of objects on a tabletop surface and converts their motion into music [20]; TinyTune is a collaborative musical instrument using

sensor motes [21]; JamMo is a mobile technological tool for music making for young children [22]; Malleable Music Making [23] is an interactive mobile system allowing multiple users to collaboratively create music over wireless networks; Beatbugs [24] are mobile instruments that can be used in collaborative networks to form large-scale collaborative compositions.

As for music listening, Camurri [25] proposed an early pioneering system where the user full-body rhythmic movements were analysed in real-time and compared with the beat of a song (extracted from the MIDI music signal). In the game Ghost In The Cave [26], two groups of users compete to collaboratively communicate emotions by full-body movement and singing voice, in order to achieve the goals of the game. Leman et al. [27] reworked the concept of social music game: the movement beat of multiple users was extracted and compared with the beat of the music the users were listening to. Users could compete among them or collaborate to win the game. Stokholm and Pasquier [28] implemented a system mixing audio representations of the mood of several users to increase collaboration and empathy. Vinyes and colleagues developed the Audio Explorer system, enabling users to concurrently modify the audio mixing of a piece of music downloaded from the Web and to share the resulting content [29]. MoodifierLive [30] is a mobile phone application that allows up to 4 users to collaboratively influence the expressive play back of MIDI content. Users move their mobile phones by taking control of one of the four music parameters: Overall Tempo, Overall Sound level, Overall Articulation, and Phrasing.

3. Concept

The SIEMPRE Podium Performance is a music performance involving two teams of three performers. These are non-expert persons rather than, e.g., experienced musicians. The performance is a kind of competitive jukebox where each team competes with the other one with the goal of making themselves and the whole audience listening to the music piece the team likes. Performers express their choices by shaking rhythmically their smartphones.

The performance starts with the smartphones laid down on a table. No music is played. The six performers are on the stage, each team occupying one side of it (see Figure 1). One performer takes one smartphone and starts shaking it. A percussive sound is reproduced, following the tempo of the movement. After a while, the other performers take their smartphone and start shaking them. Further percussive sounds are played back following the tempo of each performers. Percussive sound are different for each performer, enabling performers and the audience to recognise who is producing each sound.

Figure 1. The SIEMPRE Podium Performance presented at the Art & ICT Exhibition that took place in Vilnius (LI) in November 2013. Two teams of three performers are interacting by shaking their mobile phones: on the screen some technical data about participants is shown; team 2 is exhibiting higher cohesion and synchronisation, thus the music piece retrieved by team 2 will emerge over the one selected by team 1.

As soon as two performers in one team synchronise with each other, i.e., they shake their smartphone at almost the same tempo, a song is selected. The selected song depends on the expressive qualities of the movements of the pair, namely: (i) the tempo of their movements, and (ii) the current position of their movement in an expressive space, the energy-fluentness space. The song exhibiting the most similar characteristics is selected from a database. However, since the team is not fully collaborating (one performer is not synchronised with the other two), they listen just to the rhythmic accompaniment of the song, rather than to the whole piece. When the pair or one of its component exhibits enough leadership to involve the third team member so that all the three performers synchronise to the same tempo, the song emerges with all its voices. The other team has now to take the leadership of the performance to make its preferred music emerge. To do so, team members need to synchronise at a given tempo and they have to do it better than the other team. When the winning team changes its tempo, energy, and/or fluentness, or the other team takes the leadership of the performance, a new song emerges and is, again, modulated by the winner team.

The performance music database consists of a collection of European traditional folk music, from Portuguese fado to Lithuanian sutartine (see Section 7).

3.1. Collaboration paradigm evaluation

In *Sync4All*, a reduced version of the performance described in [31], a single team of four people has to try to completely synchronise to explore a music piece. In this version of the performance, a single music instrument is associated to each pair of performers in the team: the more they synchronise their movements, the more the listening experience is enriched by adding different instruments and voices.

Sync4All was presented and evaluated during the public exhibition "Festival della Scienza" hosted in Genoa (Italy) on November 2010. The evaluation was carried out via questionnaires. The analysis of the colected data was aimed at finding out which is the overall attitude of the users to the application, and how the participants perceived the level of collaboration and music embodiment.

Seventy-two individuals tested the application as volunteers. The questionnaire was filled up by 70 participants (38 male and 32 female) coming both from European (60 people from Italy, 4 from France, 1 from Serbia) and extra-European countries (1 from Algeria, 1 from Brazil, and 1 from Ecuador). Mean age was 22.9y (std=14.1y). Participants were clustered in basic and advanced users, depending on their skills in using mobile phones daily.

Results showed that both basic and advanced users (i) have a very positive attitude to Sync4All, (ii) perceived the level collaboration in the group and (iii) were aware that music changed according to it. Free comments and suggestions further confirmed that the participants' opinion on Sync4All was very satisfactory. One user suggested to implement a version of Sync4All in which the music tempo is aligned with the rhythm of the team: this feature is part of the current performance set-up. Another user argued that she felt constrained by the gesture to be performed in the application.

4. System set-up

Figure 2 sketches the system set-up. It encompasses several components: 6 *mobile devices* acting as clients, a laptop running EyesWeb XMI[2] acting as a server, and a *music database*.

Mobile devices capture acceleration, perform pre-processing, and send it via network to the EyesWeb server. The server carries out expressive and social features extraction from such data. Features are used to select a song for each team and to determine the winning team. Both expressive and social features are used to perform audio mixing between the two songs.

Details are provided in the following sections. To make the description concrete, components are

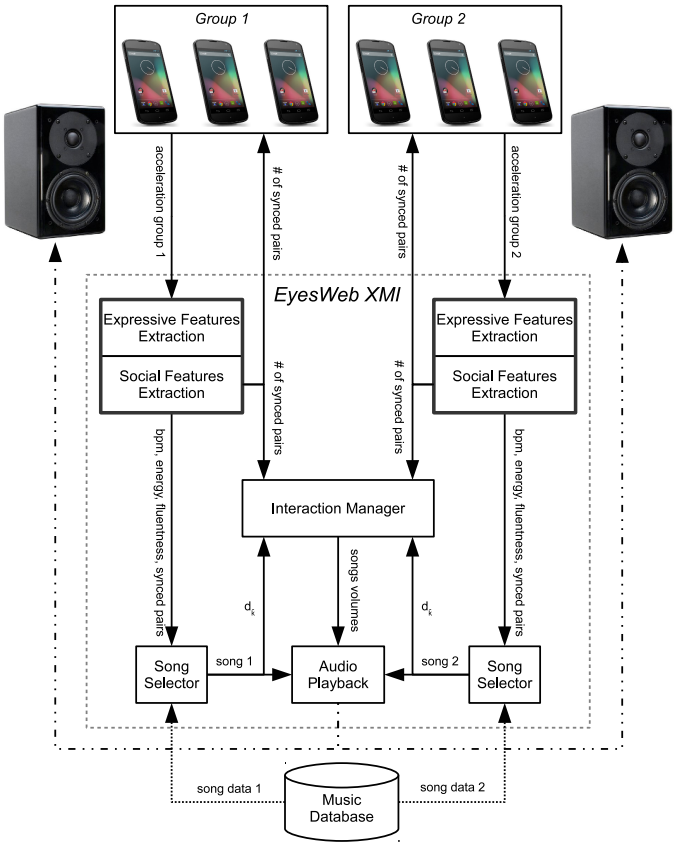

Figure 2. System set-up.

discussed with reference to the techniques used to carry out the SIEMPRE Podium Performance presented in Section 3. However, the system can be generalised to implement further instances of expressive and social retrieval and active listening applications. For example, the system can be extended to support active experience of audiovisual cultural content, e.g., for museums or consumer cultural applications.

5. Mobile devices

Mobile devices are end-user devices (e.g., Android phones) equipped with sensors including accelerometers, gyroscopes, cameras, GPS, and so on. They are connected to a wireless local area network through UDP. Mobiles run a dedicated Android application to filter out gravity acceleration and to compute the magnitude of the 3D acceleration:

$$A = \frac{\sqrt{A_x^2 + A_y^2 + A_z^2} - g}{A_{MAX}}; \quad (1)$$

where: A_x, A_y, A_z are the components of the detected acceleration along the three axes, and g is the magnitude of gravity acceleration. The acceleration A is sent to EyesWeb XMI through UDP at 50 fps.

[2] http://www.infomus.org

As depicted in 2, the *Social Features Extraction* module transmits the number of synchronised pairs to the mobile devices. Even if the concept of synchronisation is widely debated in the literature and several kinds of synchronisation are defined, in the framework of this performance synchronisation is simply intended with respect to tempo, i.e., two performers are synchronised if they shake their mobile devices at the same tempo. If the number of synchronised pairs is zero, that is, the three mobile devices of the team are moving with different tempi, then each mobile device plays a short percussive beat sound synchronised with the acceleration peaks corresponding to the user's movement strokes.

6. Expressive and social feature extraction

Expressive and social feature extraction has been implemented in the EyesWeb XMI platform and in particular in the EyesWeb Expressive Gesture Processing Library [32]. EyesWeb XMI is a software platform that allows developers to implement software modules for automatic analysis of user's expressive movement in an intuitive, visual way.

In the performance, EyesWeb XMI receives acceleration data from the mobile devices and manages the extraction of expressive and social features and the formulation of queries for search and retrieval of songs. Further, it controls the active experience of the music content, i.e., it modulates the audio mixing. The following describes the core feature extraction modules: the *Expressive Features Extraction* and the *Social Features Extraction* modules.

6.1. Expressive Features Extraction

When considering creative activities such as music playing, expressivity of human behaviour is a key component to analyse. Research works such as [33], [34] and [35] have shown that detecting and quantifying speed, acceleration, energy, smoothness, or impulsivity of one's movement may inform about emotional states.

In the SIEMPRE Podium Performance expressive feature extraction follows the multi-layered framework for analysis of expressive gesture developed by Camurri and colleagues [32]. In particular, the performance exploits the following features which are extracted for each team: (i) the tempo of the movements, (ii) the position in the energy-fluentness expressive space. Tempo was selected in order to make the performers and the audience immediately and macroscopically perceive that movement controls music selection, slow pieces being associated with low movement tempo, and quick pieces being associated with high movement tempo. Indeed, the relationship between music and movement tempo is one of the most direct and intuitive even for non-expert participants (consider for example foot tapping in time with music). The position in the energy-fluentness space is used to distinguish subtler expressive nuances between movements sharing the same tempo. Such nuances are reflected in the expressive content of the selected song. The energy-fluentness space was selected since it enables cross-modal mappings of movement features into music features. It is related to kinaesthetic, or energy-velocity spaces successfully used as control spaces to analyse and synthesise music performances, and in particular to the energy-articulation space [36]. The energy-fluentness space was already used for *ExpressiveHiFi* [37], a system where, besides the usual knobs for volume, balance, and so forth, users are also provided with additional knobs to control by their movement the expressive content of the music they are listening to.

Expressive features are extracted starting from the acceleration data mobile devices capture. The algorithms for automatic extraction of tempo, energy, and fluentness have been implemented using the EyesWeb Expressive Gesture Library [38]:

- *Tempo* (*bpm*): movement *beats per minute* (*bpm*) is computed by the FFT of the 3D acceleration magnitude A. The FFT amplitude vs. samples spectrum is analysed to find the harmonics exhibiting the maximum (A_1) and second highest (A_2) amplitudes. The first one's period (in samples) is converted to *bpm* as it follows:

$$bpm = samples * (fps/L) * 60; \tag{2}$$

  where L is the length of the buffer on which the FFT is computed. Then, *bpm* is valid if and only if:

  - the mobile device has been moved in the last 2 seconds, that is, movement energy ME is higher than a threshold;
  - $bpm > 40$;
  - the current *bpm* value has been maintained constant for the last 2 seconds;
  - the *bpm* confidence level C is higher than a threshold, where:

$$C = |A_1 - A_2|/A_1; \tag{3}$$

    That is, the difference between the highest peak amplitude (A_1) and the second highest peak (A_2) amplitude normalised by the highest peak amplitude (A_1). Figure 3 shows an example of high and low confidence levels.

- *Movement Energy* (ME): acceleration data captured by the mobile device's accelerometer as

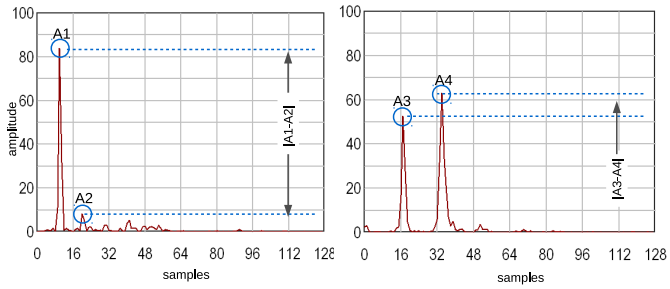

Figure 3. Example of bpm extraction and confidence level computation using FFT: (graph on the left) the difference between the highest peak amplitude (A_1) and the second highest peak amplitude (A_2) normalised by the highest peak amplitude (A_1) tends to 1; (graph on the right) the amplitudes are close, that is, the normalised difference tends to zero.

described in Section 5 is converted to Movement Energy by integrating acceleration over a time window $[t_1, t_2]$:

$$ME = \int_{t_1}^{t_2} A dt; \tag{4}$$

- *Fluentness* (Fl): the multiplicative inverse of acceleration variance over a time window $[t_1, t_2]$ is computed:

$$Fl = 1/\sigma_A; A \in [t_1, t_2]; \tag{5}$$

6.2. Social Features Extraction

This module checks whether a pair of performers in a team is tuned to the same tempo, i.e., whether they shake their mobile devices approximately with the same number of bpm. This feature is regarded as an indicator of the degree of synchronisation between a pair of performers in a team. Especially in the field of music and dance, synchronisation is used to measure the level of collaboration and entitativity of a group of people (that is, their tendency to act as a single entity) performing a joint activity [39][31].

In the proposed system, the synchronisation between a pair of performers is computed starting from the bpm they are expressing by moving their mobile devices. All the possible pairs of performers in a team are considered. In the SIEMPRE Podium Performance teams are composed by three performers, that is, there exist three pairs of performers. Synchronisation between two performers is defined as:

$$T_{u1,u2} = \begin{cases} 1, & \text{if } |bpm_{u1} - bpm_{u2}| < 15 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The tolerance of 15 bpm has been chosen according to the music database content described in 7: it corresponds to the minimum step between songs in the database. If two performers exhibit periodic movements with the same tempo (with a tolerance of 15 bpm) but shifted in phase (e.g., opposite phases, that is, 180 degrees) they will considered as synchronised. If, instead, two performers exhibit a bpm scaled by an a factor (e.g., 40 vs 80 bpm) they will be considered not in sync. The total number of synchronised pairs for each team is then sent to the *Interaction Manager* module and to the mobile devices. Based on that, audio effects (beat sounds) or songs are played back to the performers and the audience.

7. Music database, song selection and playback

The music database used in the performance consists of 46 pieces of European traditional folk music and related annotations. Songs are organised according to their tempo. This ranges from about 40 bpm, corresponding to slow pieces (*lento*), to about 200 bpm, corresponding to extremely fast pieces (*presto*). The minimum step between songs is 15 bpm. Songs are chosen so that they are clearly characterised by their tempo, i.e., tempo is quite steady and pieces do not encompass significant increase (*accelerando*) or decrease (*ritardando*) of tempo. Slow tempo pieces include for example some Polish folk music and some Romanian dances, extremely fast pieces include for example Scottish and Spanish dances. All songs are polyphonic, i.e., they include multiple musical voices (musical instruments).

Each song in the database is further annotated with respect to energy and fluentness. Both of them are quantised on three levels: *Low*, *Medium*, and *High*. Annotations for tempo, energy, and fluentness were performed by three music experts. Annotations are stored in metadata according to a tree-based hierarchical structure organised on three layers (see Figure 4):

- at the *top layer (root)*, one text file in the root folder of the database contains in each row a pair where the first item is a tempo in bpm and the second item is the path of a text file at the intermediate layer.
- at the *intermediate layer*, one text file for each tempo contains in each row a record of three items, where the first and second items are annotated values of energy and fluentness respectively (0 for *Low*, 0.5 for *Medium*, and 1 for *High*) and the third item is the path of a text file at the bottom layer.
- at the *bottom layer (leaves)*, one text file for any available combination of tempo, energy, and

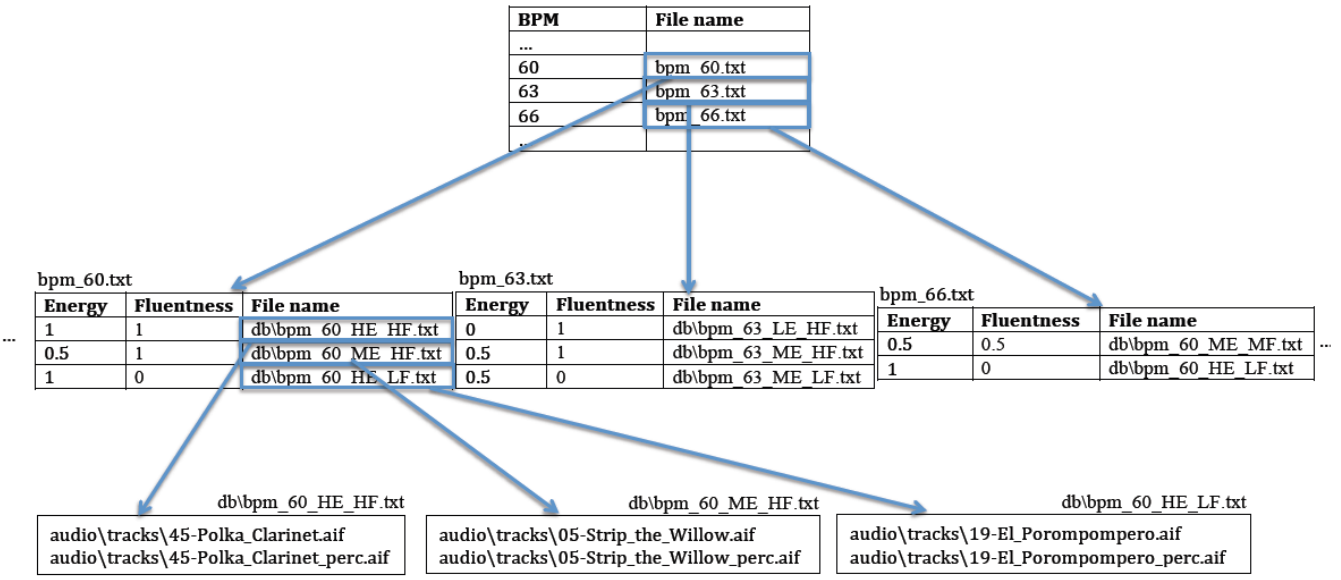

Figure 4. The tree structure for the metadata of the music database. At the root level, each tempo (measured in bpm) is associated with a file including annotations for energy and fluentness. Each combination of those is linked to a leaf including the list of music pieces exhibiting such combination of characteristics. This structure enables retrieving the result of a query by performing a visit of the tree in constant time, i.e., always requiring three steps. The files do not contain all the possible bpm values and all the possible combinations of energy and fluentness levels, but only those for which a song is available in the database.

fluentness contains the paths of the audio files of the songs exhibiting such combination of characteristics.

If on the one hand, such a multi-layered structure for metadata may require to store a possibly high number of text files, on the other hand it is optimised for enabling a fast retrieval of the selected song. Indeed, the result of a query by the *Song Selector* is retrieved by performing a visit of the metadata tree from the root to the leaf containing the songs with the desired characteristics. The visit is performed in constant time, i.e., it always needs three steps: given the desired tempo, the text file at the root layer is accessed and the path of the file including the annotations for energy and fluentness is obtained. From such a file, depending on the required level of energy and fluentness, the path of the text file including the list of available songs with those characteristics is retrieved. Finally, one song in the list is randomly selected for the performance.

It should be noted (see Figure 4) that the root file does not include all the bpm values in the range from 40 to 200, but only the tempos for which songs are available in the database. Accordingly, the intermediate layer files do not necessarily include all the eight possible combinations of *Low*, *Medium*, and *High* energy and fluentness, but only those for which songs are available in the database. In case the required tempo is not available in the database, the nearest available bmp value is selected. Similarly, the piece at the shortest distance with respect to the extracted values of energy and fluentness is retrieved (see Section 7.1).

As the file names in the leaves of the tree shows, two audio files are associated to each song. The first one is the polyphonic song itself. The second one is a rhythmic accompaniment composed for the purpose. The rhythmic accompaniment helps performers to synchronise with the rhythm of the song and can be seamlessly integrated in the song when the two components are mixed together.

7.1. Song Selector

The system set-up includes two *Song Selectors*, one for each team. Each of them receives the movement features extracted from each performer in one team and the number of synchronised pairs in the team, and selects one song for that team. The selected song is the song retrieved using the movement features of the winner performer in the team. The winner performer is identified as follows:

- for each performer $k = 1, 2, 3$, the minimum distance d_k between the tempo of the movement of the kth performer bpm_k and the tempo of each song bpm_i available in the music database is computed as:

$$d_k = \min_{i=1...N} d(bpm_k, bpm_i)\ k = 1, 2, 3; \quad (7)$$

where N is the number of songs in the database and d is the Euclidean distance:

$$d(bpm_k, bpm_i) = \sqrt{(bpm_k - bpm_i)^2}. \quad (8)$$

- the winner performer $\hat{k}$ is identified as the performer belonging to a synchronised pair having the lowest minimum distance $d_{\hat{k}}$:

$$\begin{aligned} \hat{k} &= \arg\min\{d_1, d_2, d_3\}; \\ &\text{s.t. } \exists p_j \in P : \hat{k} \in p_j \end{aligned} \quad (9)$$

where $p_j = (a, b)$ is the pair of performers a and b, $a, b = 1, 2, 3$, and P is the set of currently synchronised pairs. This means that: if performers are not synchronised, no winner is identified and no song selected; if two performers out of three are synchronised, the winner is the performer exhibiting the lowest minimum distance between the two synchronised performers; if all three performers are synchronised, the performers exhibiting the lowest minimum distance is the winner.

Once the winner performer of the team is identified, the selected song for her team is retrieved as follows:

- the set S of songs in the database whose tempo is at the minimum distance from the tempo of the winner $\hat{k}$ is computed (more than one song could have the same minimum distance):

$$S = \{\hat{i} = \arg\min_{i=1...N} d(bpm_{\hat{k}}, bpm_i)\} \quad (10)$$

- the selected song $\hat{s}$ is the song in S whose annotated energy and fluentness are at minimum distance with respect to the energy and fluentness of the movement of the winner $\hat{k}$, that is:

$$\hat{s} = \arg\min_{s \in S} d(x_{\hat{k}}, x_s)\} \quad (11)$$

where $x_{\hat{k}} = [ME_k, Fl_k]$ is the vector whose components are the energy and the fluentness of the motion of the winner performer $\hat{k}$, $x_s = [ME_s, Fl_s]$ is the annotated energy and fluentness of the song $s \in S$, and d is the Euclidean distance.

The selected song $\hat{s}$ and the distance $d_{\hat{k}}$ are sent to the *Interaction Manager* described in Section 7.2

7.2. Interaction Manager and Audio Playback

The input of the *Interaction Manager* module consists of: each team's number of synchronised pairs and distance $d_{\hat{k}}$. For the sake of clarity, $d_{1\hat{k}}$ is the winner's distance for team 1 and $d_{2\hat{k}}$ is the same distance for team 2, see Section 7.1 for details about distance computation. Its output consists of a vector:

$$V = [V_{p1}, V_{r1}, V_{p2}, V_{r2}] \quad (12)$$

where: (i) V_{pi} is the volume at which the **polyphonic** version of the winner song for team i has to be played at; (ii) V_{ri} is the volume at which the **rhythmic** version of the winner song for team i has to be played at. This vector is sent to the *Audio Playback* module, together with the songs associated to each team. The audio volumes are adjusted depending on the values stored in V and the final audio output is played back to the users through speakers. The content of V depends on the amount and type of intra- and inter-teams interaction as it follows:

1. *intra-team* (for each team i):
   (a) *0 synchronised pairs*: $V_{pi} = V_{ri} = 0$; i.e., in this case the mobile devices will produce beat sounds, see Section 5);
   (b) *1 or 2 synchronised pairs*: $V_{pi} = 0, V_{ri} = 1$;
   (c) *3 synchronised pairs*: $V_{pi} = 1, V_{ri} = 0$;
2. *inter-team*:
   (a) *3 synchronised pairs for both teams*: if $(d_{1\hat{k}} < d_{2\hat{k}})$ then $V_{p2} = 0$ else $V_{p1} = 0$;
   (b) *3 synchronised pairs for just one team j*: then $V_{pi} = 0, i \neq j$;

The goal of the SIEMPRE Podium Performance is to engage performers in joint music content exploration and retrieval. The music retrieved by the team exhibiting higher participants' cohesion and synchronisation will emerge: (steps 1a-1c) a higher number of synchronised pairs inside the team (higher cohesion) will increase the audio polyphony; (steps 2a and 2b) a higher similarity between the team participants and audio rhythm (synchronisation) will make the corresponding team's song emerge over the other team's one.

8. Conclusion

This paper presented the SIEMPRE Podium Performance and the system developed for it. The system includes algorithms for extraction of social and expressive features characterising natural behaviour of performers and techniques for automated retrieval of and

active listening to music. The system enables non-expert members of an audience to be involved in the show and allow them to interpret the music they like. The performance was repeated two times at the Art & ICT Exhibition 2013. It may contribute to open novel directions in both research and market niches for future User Centric Media. Next developments will focus on extraction of further social features, on query formulation, and on other kinds of capture devices (e.g., Leap-Motion). For example, our performance is percussive feedbacks inducing users to perform rhythmic gestures to interact between them: other types of feedback, e.g., a violin melody, could be introduced to trigger fluent movements in the performers requiring the extraction of different expressive and social features. By adding new gestures types, users could implicitly choose their audio instrument and interpret the output melody, like in improvisational music. Results from the research described in this paper include new developments to the EyesWeb software platform, to the library for real-time analysis of human behaviour and to libraries for real-time processing of movement and audio content on mobiles. These improvements are currently exploited in scientific experiments and artistic projects in the framework of the EU Culture Project MetaBody[3].

Acknowledgements

This research originated from the results obtained in the EU FP7 ICT FET Project SIEMPRE, and is currently partially funded by the EU Culture Project MetaBody (Agreement Number 2013-1572/001-001). We thank our colleagues Paolo Coletta, Alberto Massari, and Simone Ghisio for their precious contributes to software development. We thank Corrado Canepa and Giacomo Lepri for their contribute to the design of the music content. Finally, we thank Teresa De Martino and Beatrice Marquez-Garrido for their kind support and for the permission to use in this paper their ICT2013 Vilnius photos of our demo.

References

[1] Lee, S.W., Essl, G. and Mao, Z.M. (2014) Distributing mobile music applications for audience participation using mobile ad-hoc network (manet). In *NIME'14 (New Interfaces for Musical Expression)*.

[2] Dahl, L., Herrera, J. and C., W. (2011) Tweetdreams: Making music with the audience and the world using real-time twitter data. In *NIME'11 (New Interfaces for Musical Expression)*.

[3] Camurri, A., Volpe, G., Vinet, H., Bresin, R., Fabiani, M., Dubus, G., Maestre, E. *et al.* (2010) User-centric context-aware mobile applications for embodied music listening. In *User Centric Media* (Springer), 21–30.

[4] Marsh, K., Richardson, M. and Schmidt, R. (2009) Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science* **1**: 320–339.

[5] Bianchi-Berthouze, N., Kim, W.W. and Patel, D. (2007) Does body movement engage you more in digital game play? and why? In *Proceedings of Affective Computing and Intelligent Interaction*: 102–113.

[6] Lindley, S.E., Couteur, J.L. and Berthouze, N. (2008) Stirring up experience through movement in game play: effects on engagement and social behaviour. In *Proceedings of ACM Conference on Human Factors in Computing Systems*: 511–514.

[7] Mancas, M., Madhkour, R.B., De Beul, D., Leroy, J., Riche, N., Rybarczyk, Y.P. and Zajéga, F. (2011) Kinact: a saliency-based social game. In *Proceedings of the 7th International Summer Workshop on Multimodal Interfaces eNTERFACE11*, **8**.

[8] I. Laso-Ballesteros and P. Daras (Eds.), User Centric Future Media Internet, EU Commission, September 2008.

[9] Pentland, A. (2007) Social signal processing. *IEEE Signal Processing Magazine* **2**(4): 108–111.

[10] Hung, H., Huang, Y., Friedland, G. and Gatica-Perez, D. (2011) Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech & Language Processing* **19**(4): 847–860.

[11] Hung, H. and Gatica-Perez, D. (2010) Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* **12**(6): 563–575.

[12] Vinciarelli, A., Pantic, M., Bourland, H. and Pentland, A. (2008) Social signals, their function, and automatic analysis: a survey. In *Proceedings of the 10th International Conference on Multimodal Interfaces, ICMI08*: 61–68.

[13] Lukowicz, P., Pentland, S. and Ferscha, A. (2012) From context awareness to socially aware computing. *Pervasive Computing, IEEE* **11**(1): 32–41.

[14] Dong, W., Lepri, B. and Pentland, A. (2011) Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia* (ACM): 134–143.

[15] Jayagopi, D., Kim, T., Pentland, A. and Gatica-Perez, D. (2012) Privacy-sensitive recognition of group conversational context with sociometers. *Multimedia systems* **18**(1): 3–14.

[16] Bortz, B., Salazar, S., Jaimovich, J., Knapp, R. and Wang, G. (2012) Shemp: A mobile framework for shared emotion, music, and physiology. In *Proceedings of the Third International Workshop on Social Behaviour in Music, 14th ACM International Conference on Multimodal Interaction*.

[17] Blaine, T. and Fels, S. (2003) Collaborative musical experiences for novices. *Journal of New Music Research* **32**: 411–428.

[18] Jordà, S. (2005) Multi-user instruments: models, examples and promises. In *Proceedings of the 2005 conference on New interfaces for musical expression*, NIME '05 (Singapore, Singapore: National University of Singapore): 23–26.

[3] www.metabody.eu

[19] Jordà, S., Geiger, G., Alonso, M. and Kaltenbrunner, M. (2007) The reactable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proceedings of the 1st international conference on Tangible and embedded interaction*, TEI '07 (New York, NY, USA: ACM): 139–146.

[20] Patten, J., Recht, B. and Ishii, H. (2002) Audiopad: a tag-based interface for musical performance. In *Proceedings of the 2002 conference on New interfaces for musical expression*, NIME '02 (Singapore, Singapore: National University of Singapore): 1–6.

[21] Newman, B., Sanders, J., Hughes, R. and Jurdak, R. (2010) Tinytune, a collaborative sensor network musical instrument. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, SenSys '10 (New York, NY, USA: ACM): 375–376.

[22] Myllykoski, M. and Paananen, P. (2009) Towards new social dimensions for children's music making - jammo as a collaborative and communal m-learning environment. In *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music*, ESCOM 2009: 366–371.

[23] Tanaka, A. (2004) Malleable mobile music. In *Adjunct Proceedings of the 6th International Conference on Ubiquitous Computing (UBICOMP)*.

[24] Weinberg, G., Aimi, R. and Jennings, K. (2002) The beatbug network: a rhythmic system for interdependent group collaboration. In *Proceedings of the 2002 conference on New interfaces for musical expression* (National University of Singapore): 1–6.

[25] Camurri, A. and Ferrentino, P. (1999) Interactive environments for music and multimedia. *Multimedia Systems* **7**(1): 32–47.

[26] Lindström, E., Camurri, A., Friberg, A., Volpe, G. and Rinman, M.L. (2005) Affect, attitude and evaluation of multisensory performances. *Journal of New Music Research* **34**(1): 69–86.

[27] Leman, M., Demey, M., Lesaffre, M., van Noorden, L. and 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*.

[28] Stockholm, J. and 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*.

[29] Vinyes, M., Bonada, J. and Loscos, A. (2006) Demixing commercial music productions via human-assisted time-frequency masking. In *Proceedings of the 120th AES Convention*.

[30] Fabiani, M., Dubus, G. and Bresin, R. (2011) MoodifierLive: Interactive and collaborative expressive music performance on mobile devices. *Proceedings of NIME 2011* : 116–119.

[31] Varni, G., Mancini, M. and Volpe, G. (2012) Embodied cooperation using mobile devices: presenting and evaluating the sync4all application. In *AVI*: 312–319.

[32] Camurri, A., Mazzarino, B. and Volpe, G. (2004) Expressive interfaces. *Cognition, Technology and Work* **6**: 15–22.

[33] Wallbott, H.G. and Scherer, K.R. (1986) Cues and channels in emotion recognition. *Journal of Personality and Social Psychology* **51**(4): 690–699.

[34] Wallbott, H.G. (1998) Bodily expression of emotion. *European Journal of Social Psychology* **28**: 879–896.

[35] Gallaher, P.E. (1992) Individual differences in nonverbal behavior: Dimensions of style. *Journal of Personality and Social Psychology* **63**(1): 133–145.

[36] Canazza, S., Poli, G., RodÃă, A. and Vidolin, A. (2003) An abstract control space for communication of sensory expressive intentions in music performance. *Journal of New Music Research* **32**(3): 281–294.

[37] Camurri, A., Volpe, G., De Poli, G. and Leman, M. (2005) Communicating expressiveness and affect in multimodal interactive systems. *Multimedia, IEEE* **12**(1): 43 – 53.

[38] Camurri, A., Mazzarino, B. and Volpe, G. (2004) Analysis of expressive gesture: The eyesweb expressive gesture processing library. *Lecture notes in computer science* .

[39] Glowinski, D., Mancini, M., Cowie, R., Camurri, A., Chiorri, C. and Doherty, C. (2013) The movements made by performers in a skilled quartet: a distinctive pattern, and the function that it serves. *Frontiers in Psychology* **4**(841).