UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Article scientifique | Article | 2015

Published version | Public access

This is the published version of the publication, made available in accordance with the publisher's policy.

Expressive non-verbal interaction in a string quartet: an analysis through head movements

Glowinski, Donald; Dardard, Floriane; Gnecco, Giorgio; Piana, Stefano; Camurri, Antonio

How to cite

GLOWINSKI, Donald et al. Expressive non-verbal interaction in a string quartet: an analysis through head movements. In: Journal on Multimodal User Interfaces, 2015, vol. 9, n° 1, p. 55–68. doi: 10.1007/s12193-014-0154-3

This publication URL: https://archive-ouverte.unige.ch/unige:178957
Publication DOI: 10.1007/s12193-014-0154-3

© The author(s). This work is licensed under a Backfiles purchase (National Licenses Project)
https://www.unige.ch/biblio/aou/fr/guide/info/references/licences/

Last deposit update in Archive ouverte UNIGE on 25.07.2024 16:17

J Multimodal User Interfaces (2015) 9:55–68
DOI 10.1007/s12193-014-0154-3

ORIGINAL PAPER

Expressive non-verbal interaction in a string quartet: an analysis through head movements

Donald Glowinski · Floriane Dardard · Giorgio Gnecco · Stefano Piana · Antonio Camurri

Received: 21 January 2014 / Accepted: 21 April 2014 / Published online: 14 May 2014
© OpenInterface Association 2014

Abstract The present study investigates expressive non-verbal interaction in the musical context starting from behavioral features extracted at individual and group levels. Four groups of features are defined, which are related to head movement and direction, and may help gaining insight on the expressivity and cohesion of the performance, discriminating between different performance conditions. Then, the features are evaluated both at a global scale and at a local scale. The findings obtained from the analysis of a string quartet recorded in an ecological setting show that using these features alone or in their combination may help in distinguishing between two types of performance: (a) a *concert-like* condition, where all musicians aim at performing at best, (b) a *perturbed* one, where the 1st violinist devises alternative interpretations of the music score without discussing them with the other musicians. In the global data analysis, the discriminative power of the features is investigated through statistical tests. Then, in the local data analysis, a larger amount of data is used to exploit more sophisticated machine learning techniques to select suitable subsets of the features, which are then used to train an *SVM* classifier to perform binary classification. Interestingly, the features whose discriminative power is evaluated as large (respectively, small) in the global analysis are also evaluated in a similar way in the local analysis. When used together, the 22 features that have been defined in the paper demonstrate to be efficient for classification, leading to a percentage of about 90 % successfully classified examples among the ones not used in the training phase. Similar results are obtained considering only a subset of 15 features.

Keywords Automated analysis of non-verbal behavior · Head ancillary gestures · Focus of attention · Feature selection · Support vector machines

D. Glowinski (✉)
Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland
e-mail: donald.glowinski@unige.ch

F. Dardard
Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, Paris, France
e-mail: floriane.dardard@telecom-paristech.fr

G. Gnecco
IMT, Institute for Advanced Studies, Lucca, Italy
e-mail: giorgio.gnecco@imtlucca.it

S. Piana · A. Camurri
Casa Paganini-InfoMus Research Centre, DIBRIS, University of Genoa, Genoa, Italy
e-mail: stefano.piana@dist.unige.it

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

1 Introduction

Among human interactive and social activities, performing music is a well-known case in which non-verbal communication plays a fundamental role. Several studies have used observational and interview methods to explore the way musicians interact and determine the overall quality of experience [1]. Others, including the study reported here, investigate the interaction by means of quantitative measures, with a particular focus on expressive alignment processes in communication. The literature on alignment grows out of linguistics research on convergence between speakers, but it has broadened to include various nonverbal behaviors (e.g., [2] reviews studies on speech prosody, turn taking, joint attention, backchanneling, head nods, smiles and mirroring/contagion effects). Within this framework, string quartets (SQs) have been identified as a particularly promising

context for investigating expressive and adaptive interactions of groups of people. The SQ scenario involves a particular social structure, which one would expect to be reflected in a particular style of communication [3]. In a SQ, all the musicians contribute equally to the performance of the group. There is some degree of leadership, usually played by the 1st violinist, but not the kind of hierarchy that can be seen in an orchestra (conductor or concertmaster *vs.* other musicians). In this perspective, the SQ has been described as a *self-managed team*, i.e., a working structure where all partners share roughly equal responsibility in the development of a common project [4].

In this context, the aim of the present study (which is an extended version of [5], presented at IEEE ACII 2013) is to shed light on the way musicians behave in performing the co-operative and emotionally-engaged task of playing in a SQ: in particular, we want to measure how the expressive behavior of the musicians may change as a consequence of modifications in the performing conditions. In the measurement process, we focus on visual features, related to the movements of the players. Compared to [5], in Sect. 5 of the present work we have used more sophisticated machine learning techniques, which have the advantage of being able to consider more features at the same time. On the other hand, such techniques also need more data for their training, as the associated model complexity is larger. For this reason, in Sect. 5 more punctual features have been used, thus enlarging the number of available training data, and making possible the use of machine learning techniques such as support vector machines (*SVM*s). Moreover, due to the use of short time windows, the number of samples considered in the local analysis performed in Sect. 5 is substantially larger than the one used in global analysis performed in [5] and refined in Sect. 4 of the present work.

Several works have already shown how the movements of a player can carry information about the performance of a musical piece (e.g., by conveying different expressive intentions). In this paper, we focus on head movements, which belong to the wider category of *ancillary* or *accompanist gestures* [6], i.e., movements of the body of a music player or of a music instrument which are not directly related to the production of the sound (in contrast to *instrumental* or *effective gestures*, which are directly involved in sound production). Besides head movements, other ancillary gestures are, e.g, the movements of the hands of a harpist during and after string plucking [7]. The movements of the bell of a clarinet are often classified as ancillary gestures [8] since they are performed spontaneously by the music player, although they have actually a direct role in the production of sound, since they are movements of a sound source (the clarinet). Finally, the movements of the bows of string players during a performance are (mainly) instrumental gestures.

Instrumental gestures are obviously informative since, without them, musicians would not be able to express the different musical ideas they want to communicate. Ancillary gestures are informative, too, since they often allow one to recognize different expressive intentions, without looking at the instrumental gestures/listening to the performance. For instance, for the case of a piano player, Davidson investigated how visual information alone is often sufficient to discriminate among performances of the same piece of music played with different expressive intentions (inexpressive, normal and exaggerated) [9], and found that the larger the amplitude of the movement, the deeper the expressive intention [10]. This finding was also confirmed by other studies, e.g., Castellano et al. investigated the discriminatory power of several movement-related features (again, for the case of a piano player) [11], whereas Palmer et al. showed how the movement made by the bell of a clarinet is larger when the player performs more expressive interpretations of the same piece [12]. However, these works focus on a performance by one player only. More recent studies address non-verbal communication in larger musical ensembles such as a string quartet [13] and a section of an orchestra [14,15]. Interpersonal interaction among musicians was also studied in [16–21], whereas [22–24] focused in conductor-musicians scenarios, and [22–27] investigated the musician–listener interaction.

Among ancillary gestures, as already mentioned, the focus of this paper is on head movements, which are known to play a central role in non-verbal communication in general [28] and in music in particular [15,29,30]. Head movements were investigated, e.g., in [5,31–34] to estimate the position of a common point of interest for a group of people, and in [15] to see how they depend on the presence/absence of a such a common point of interest. In principle, eye-gazes would be better suited than head directions for these applications. However, eye-gaze tracking equipment is still nowadays intrusive and costly. Moreover, previous studies have shown that head direction and eye-gaze are often correlated [31–33,35]. In [34], the contribution of each person in the determination of the position of the point of interest has been evaluated using features related to head movements, combined with cooperative game theory.

In a musical context, head movements may express the way musicians understand the phrasing and breathing of the music they are performing, and so provide information about the high-level emotional structures in terms of which the players are interpreting the music. In addition, head movements are not too sensitive to score details (differently from, e.g., the movement of the bow tip in the case of a string player). Additional information indicating how each musician stands with respect to the group as the performance unfolds, may be obtained by studying the movement of musicians' heads with respect to the positions of several points of interest.

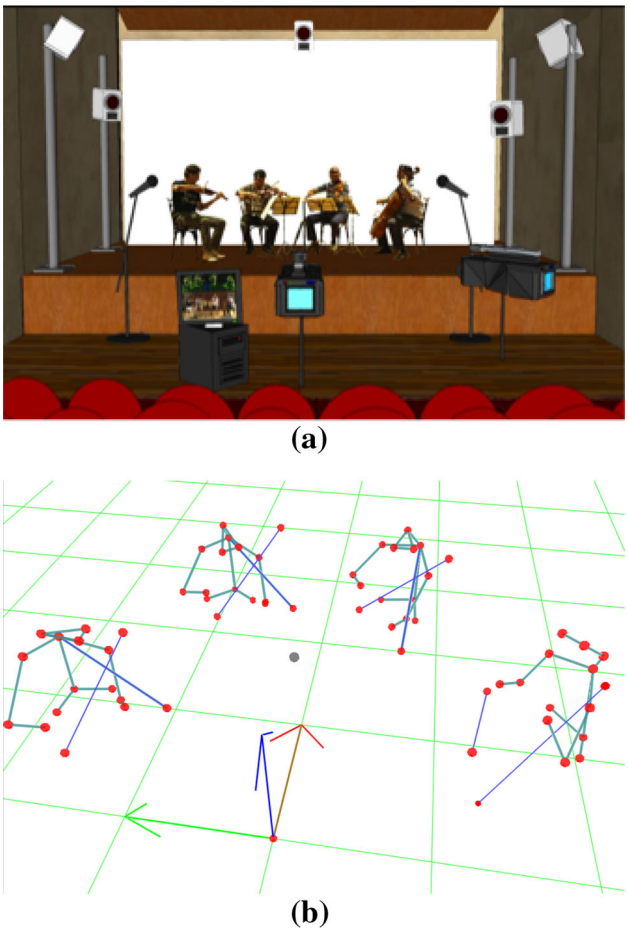


Fig. 1 A picture of the four musicians of the SQ (Quartetto di Cremona) studied in this paper (**a**), together with their motion capture (MoCap) 3D representation (**b**). The musicians are wearing MoCap reflective markers and physiological and acoustic sensors. The subjective center of the SQ, the *ear*, is represented by a *black dot* and corresponds to a reflective marker mounted above a tripod situated among the musicians

As proposed in [21], the *ear* of the quartet is a prominent example of such points of interest. The SQ *ear* refers to a fixed subjective center of the SQ, whose position is defined by the musicians themselves, and which is located at nearly equal distance from each of them (see Fig. 1). In this work, it has been identified asking the musicians to indicate a position on the stage, then placing a tripod (and a reflective marker attached to the tripod) in such a position. The SQ *ear* is so called because it refers to the best location of an imaginary listener who would receive the musical contributions of all the musicians. This center is expected to function as a reference point for all the musicians during the performance and to help them in coordinating and achieving a coherent sound.

In this direction and following [32], in the present work, four groups of features have been implemented to evaluate:

1. how the heads' directions of the four musicians converge toward the SQ *ear*;
2. how much the musicians move jointly forward and backward with respect to the SQ *ear*;
3. how the heads' directions of each subset of three musicians converge toward the head of the remaining one;
4. how much the head of each musician is directed toward each other musician.

Hence, a different movement behavior of the group with respect to the *ear* or other points of interest may be expected to reflect different expressive performing conditions.

The paper is organized as follows: Sect. 2 describes the multimodal setup and the experimental procedure, Sect. 3 details the behavioral features implemented to characterize group and individual expressive performance, Sect. 4 presents the results obtained when examining the single features, and Sect. 5 refers to classification results obtained using their combination. Finally, Sect. 6 discusses the main findings and presents some conclusions.

2 Subjects and stimuli

2.1 Choice of professional concert level musicians

The Quartetto di Cremona, an internationally recognized string quartet, was invited to participate to the experiment. Preliminary encounters confirmed that the components of this quartet show key qualities that made them suitable for conducting this study. They were able to tolerate disturbance created by the multimodal setup (videocameras, markers, and on-body sensors) thanks to their longstanding experience of performance in a variety of environmental situations (concert hall, television and radio broadcastings). They understood and replied in detail to the experimenters' demands as they are accustomed to working collaboratively with life contemporary composers for whom they created artworks. They demonstrated high flexibility in performing a variety of styles and have developed well-advanced strategies to rehearse altogether. The piece that was selected is part of their repertoire.

2.2 Choice of the musical fragment

The music piece performed by the SQ during the experiment was extracted from the Allegro of the String Quartet No. 14 in D minor, known as Death and the Maiden, by Franz Schubert. This piece is a staple of the quartet's repertoire and stirs together a number of very contrasted musical elements including homorhythmic structure where musicians tend to play at unison, fugato writing styles which replicate the musical subject over the different instruments or concerto style melodic development interpreted by the 1st violinist and accompanied by repetitive chords and tremolos of the other musicians.

2.3 Procedure

Two sessions of recordings were done with the Quartetto di Cremona (July, 13th and 14th, 2011) following two experimental procedures. In the first procedure (*condition A*, experimented on the 1st day), the four musicians were instructed to play five times the Schubert music piece at best in a concert-like situation. In the second protocol (*condition B*, experimented on the 2nd day), the 1st violinist of the string quartet devised alternative interpretations of the music score, which contradict the usual interpretation (e.g., playing forte where nuance is written piano, speeding up when a rallentando is requested, etc.). Also this procedure was repeated five times. The other members of the quartet were not aware of these new versions before playing. For each recording, the quality of each performance was assessed by the musicians themselves through post-performance ratings on a 7-items Likert scale (e.g., *expressivity* and *group cohesion* were evaluated asking and answering questions such as "how emotionally engaging was your performance?" and "how did you manage to coordinate with the other musicians?", see [36]).

2.4 Apparatus and set-up

The experiment was made within the EU Project SIEMPRE (Social Interaction Entertainment Using Music Performance)[1] and took place at Casa Paganini - InfoMus Research Centre in Genova, Italy,[2] in a 250-seat auditorium, an environment similar to a concert hall, suitable for experiments in ecological setups (see Fig. 1a). A multimodal recording platform was set up to capture and analyze the movement, audio, and physiological data of the musicians. In particular, the musicians' movement behavior was captured by means of the Qualisys Motion Capture system,[3] equipped with 7 cameras. 16 reflective markers were placed on each musician's joints and 3 other markers were located on each instrument. In particular, 3 reflective markers were placed on each musician's head (1 marker on the back of the neck, 2 markers above the eyes). The positions of the markers were extracted by the Qualisys Motion Capture system using the data collected by the seven cameras. Original real-time applications based on the EyesWeb eXtended Multimodal Interaction (XMI)[4] software platform were developed to synchronize the Qualisys MoCap data together with the video and audio data. Before being analyzed, the MoCap data were linearly interpolated when the tracking of the markers was not possible, due to possibly undetected markers or missing labels associated with the markers. However, among the various markers, the ones located on the heads did not actually require corrections (likewise other markers, such as the ones located on the shoulders), which is one of the reasons for which such markers were used in the subsequent data analysis (another one is that they are expected to be naturally related to the direction to which the musicians are focusing their attention).

2.5 Selected data

The present paper focuses on one particular component of the recordings: the time-series data of the positions of the heads of the musicians. So, among the 76 available markers, only 12 markers were exploited in the data analysis. Moreover, in a similar way as [15,21,30,34], only their coordinates in the horizontal plane were used. The code we developed to perform the data analysis of such time series was written in MATLAB.

Table 1 summarizes the data used in the analysis, showing the number (5) of recordings available for each condition, the duration (of the order of 2–3 min) of each recording, the number of frames (of the order of ten thousand) available for each recording, and, for the local analysis performed in Sect. 5, the number of time windows obtained for each recording, for a data fragmentation corresponding to 2- and 4-s time windows, respectively (see Sect. 5 for more details on their definitions). Moreover, the frame rate of the cameras was 100 frames per second, the maximum number of markers detected in a frame was 76, and the number of markers per frames actually used in the data analysis was 12. Here, one can notice that the larger variability of the duration of the recordings in condition *B* is due to the presence of sudden unexpected accelerandos/rallentandos played by the 1st violinist under such a condition.

Table 1 For each recording: its duration, its number of frames, and, for the local analysis performed in Sect. 5, its number of time windows, for both cases of 2- and 4-s time windows

Rec. (cond.)	Time length	# Frames	# 2-s windows	# 4-s windows
1(*A*)	2′19″	13,919	138	68
2(*A*)	2′35″	15,506	154	76
3(*A*)	2′29″	14,968	148	73
4(*A*)	2′29″	14,962	148	73
5(*A*)	2′27″	14,736	146	72
1(*B*)	2′25″	14,548	144	71
2(*B*)	2′10″	13,048	129	64
3(*B*)	2′51″	17,142	170	84
4(*B*)	3′28″	20,845	207	103
5(*B*)	2′32″	15,246	151	75

For all cases, the frame rate of the cameras was 100 frames per second, the maximum number of markers detected in a frame was 76, and the number of markers per frames actually used in the data analysis was 12

[1] http://www.siempre.infomus.org.

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

[3] http://www.qualisys.com.

[4] http://www.infomus.org/eyesweb_eng.php.

In Sect. 3, several static and dynamic behavioral features of the SQ are described. They are related to the movement of musicians' heads with respect to specific points of interest (e.g., the SQ *ear*).

3 Description of the implemented behavioral features

This section details the features implemented to characterize group expressive behavior and to distinguish between the two performing conditions A and B. The music players are numbered from 1 to 4, so the number 1, 2, 3, 4 denotes, respectively, the 1st violinist, the 2nd violinist, the violist, and the cellist. The frames of each recording are denoted by k ($k = 1, \ldots, N_{\text{frames}}$). The number of frames depends slightly on the recording, but for simplicity of notation, this dependence is not shown in the following formulas.

3.1 Convergence of the heads' directions toward the ear

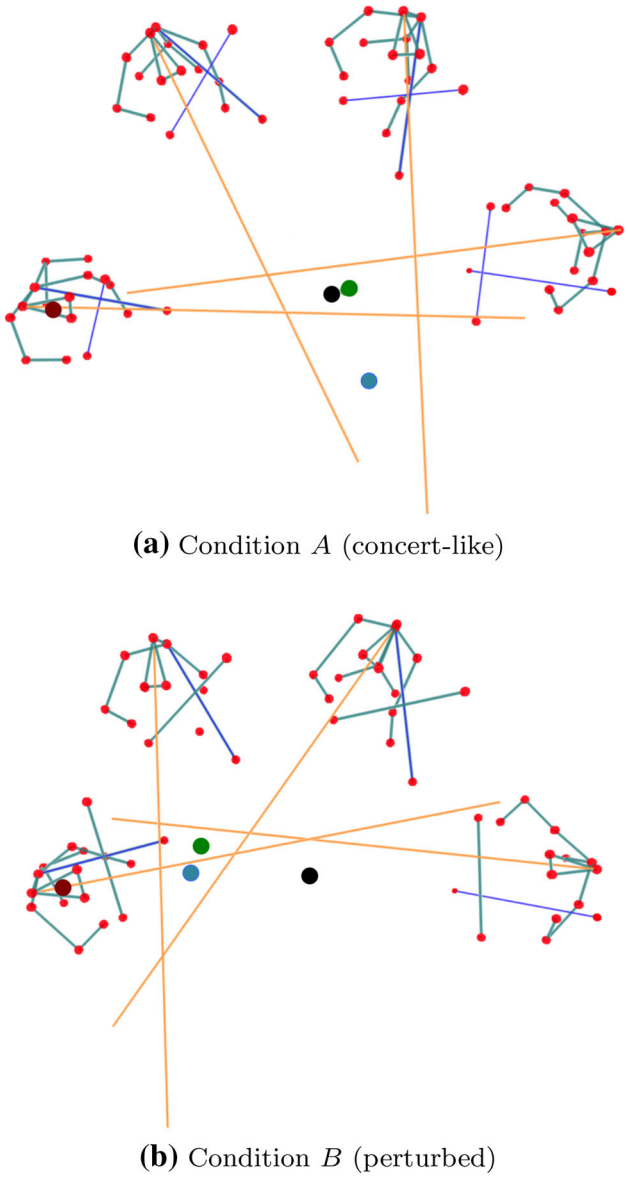
The first behavioral feature F_1 evaluates how the heads' directions of the four musicians converge toward the SQ *ear* (see Sect. 1).

The following procedure has been followed, for each frame k of each recording.

1. For each musician i ($i = 1, \ldots, 4$), compute the current position vector $\mathbf{p}_i^{(k)}$ in the horizontal plane of the musician's head center of gravity (COG) as the mean of the position vectors describing the three markers located on the musician's head. Then, define the current direction $\mathbf{d}_i^{(k)}$ in the horizontal plane of the musician i's head as the unit vector connecting the COG of his head to the point located in the middle of the line between the two other markers above his eyes (see Fig. 2).
2. For each musician i ($i = 1, \ldots, 4$), consider the half-line $\mathbf{HL}_i^{(k)}$ starting from the point $\mathbf{p}_i^{(k)}$ and with direction $\mathbf{d}_i^{(k)}$, i.e., the set of all the points with position vectors

   $$\mathbf{p}_i^{(k)} + t\mathbf{d}_i^{(k)},$$

   where $t \geq 0$ is any nonnegative real number.
3. For each pair (i, j) of musicians ($i, j = 1, \ldots, 4, i < j$), compute the position vector $\mathbf{p}_{i,j}^{(k)}$ of the intersection between the two half-lines $\mathbf{HL}_i^{(k)}$ and $\mathbf{HL}_j^{(k)}$. As $\mathbf{p}_i^{(k)} \neq \mathbf{p}_j^{(k)}$, such an intersection exists if and only if the following condition is met:
   − the algebraic linear system (in the real unknowns u and v)



(a) Condition A (concert-like)

(b) Condition B (perturbed)

Fig. 2 Illustration of features F_1 and $\mathbf{F}_3$ measuring how the heads' directions of the four musicians converge toward the *ear* and how the heads' directions of each subset of three musicians converge toward the head of the remaining one, respectively. The figure shows a snapshot of the heads' markers positions of the music players when condition A (*concert-like*) and condition B (*perturbed*) are tested, respectively. White half-lines refer to the heads' directions and the *green dot* corresponds to the position of the *point of total convergence* (PoTC) (i.e. where all musician heads are converging). The *blue point* represents the *point of partial convergence* associated with the 1st violinist (PoPC_1) (i.e. where the subset of the other 3 musician's heads are converging). Similarly-defined points are associated with the other musicians. One can observe that in the first snapshot, all musicians' head directions converge toward the *ear* of the SQ (*black dot* in the picture above), whereas in the second snapshot, the 2nd violinist, violist and cellist heads' directions converge toward the 1st violinist's one (*brown dot*). The positions of all these points are used to compute the features F_1 and $\mathbf{F}_3$, see formulas (1) and (3)

$$\mathbf{p}_i^{(k)} + u\mathbf{d}_i^{(k)} = \mathbf{p}_j^{(k)} + v\mathbf{d}_j^{(k)}$$

has a unique solution (this happens if and only if $\mathbf{d}_i^{(k)}$ is not parallel to $\mathbf{d}_j^{(k)}$), and both the obtained u and v are nonnegative.
When the condition above holds, the position vector $\mathbf{p}_{i,j}^{(k)}$ is then defined equivalently as

$$\mathbf{p}_{i,j}^{(k)} = \mathbf{p}_i^{(k)} + u\mathbf{d}_i^{(k)},$$

or

$$\mathbf{p}_{i,j}^{(k)} = \mathbf{p}_j^{(k)} + v\mathbf{d}_j^{(k)}.$$

The procedure is repeated 6 times, determining—for the frames for which they exist—the 6 position vectors $\mathbf{p}_{1,2}^{(k)}$, $\mathbf{p}_{1,3}^{(k)}$, $\mathbf{p}_{1,4}^{(k)}$, $\mathbf{p}_{2,3}^{(k)}$, $\mathbf{p}_{2,4}^{(k)}$, $\mathbf{p}_{3,4}^{(k)}$ of the 6 pairwise intersections.

4. Denote by $I^{(k)}$ the subset of the pairs (i, j) of musicians $(i, j = 1, \ldots, 4, i < j)$ for which the pairwise intersections above exist at frame k, and by $|I^{(k)}|$ its cardinality. If $I^{(k)}$ is nonempty, then the position vector of the *point of total convergence* (PoTC)—the point where the directions of all musician's head directions converge (see Fig. 2)—is defined as

$$\mathbf{p}_{PoTC}^{(k)} = \frac{\sum_{(i,j)\in I^{(k)}} \mathbf{p}_{i,j}^{(k)}}{|I^{(k)}|}.$$

If $I^{(k)}$ is empty, the PoTC is not defined at frame k.

5. Denote by $\mathbf{e}$ the fixed position vector of the SQ *ear* and evaluate the distance $\|\mathbf{p}_{PoTC}^{(k)} - \mathbf{e}\|$ between PoTC and *ear*. When the PoTC is not defined, the distance is set equal to its maximum value achieved in the frames of the recording for which the PoTC exists.

The first behavioral feature F_1 is defined as the median distance between the PoTC and the *ear*:

$$F_1 = \text{median of } \|\mathbf{p}_{PoTC}^{(k)} - \mathbf{e}\|. \quad (1)$$

The median is computed with respect to the frames of each single recording, and is less sensitive to outliers than the mean[5] (an alternative to the median could be a "trimmed" mean, which excludes outliers). Indeed, even in the case of a recording for which the parallelism condition described above never occurs, still one reason for the presence of outliers is an "almost parallelism" condition for the head directions of two different players, which in a small percentage of frames may affect severely the determination of the positions of the point of total convergence (the same remark holds also for the points of partial convergence, defined later in Sect. 3.3). Another motivation is that, in some frames, some intersections between the players' head directions used to define the points of total and partial convergence could not exist (so, an average with respect to the remaining intersections is performed). As a final step, the feature F_1 defined in formula (1) is also normalized, dividing it by its median over all the recordings, in such a way that its range is comparable to the ones of the other features. The results of the data analysis in Sect. 4.1 refer to such a normalized feature (Fig. 3).

[5] For what concerns the definitions of some features, we have corrected a typo present in [5], which reported an old definition of the features F_1 and $\mathbf{F}_3$ (given in terms of means over the frames instead than medians over the frames, as done in the present manuscript), although its numerical results about such features were actually obtained according to the same definitions of the present manuscript.

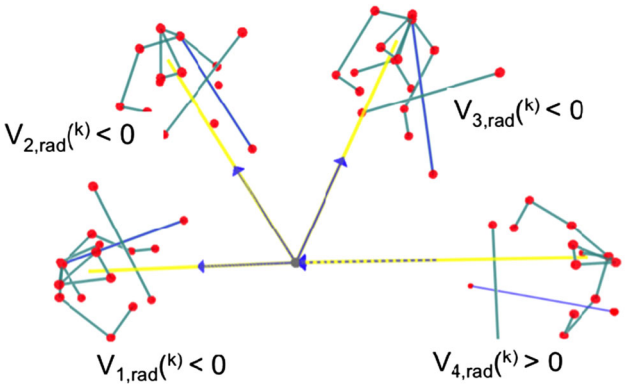


Fig. 3 Illustration of feature F_2 measuring how much the musicians move jointly forward and backward with respect to the *ear*. The figure shows an example in which cohesion is not maximal as the radial components of the head velocities of the 1st violinist, the 2nd violinist and the violist are smaller than 0 (their heads are moving away from the *ear*), but the radial component of the cellist's head velocity is greater than 0 (it is moving toward the *ear*). Such radial components are used to compute the feature F_2 (see formula (2))

3.2 Joint movement dynamics of the heads toward the ear

The second behavioral feature F_2 evaluates how much the musicians move *jointly* forward and backward with respect to the SQ *ear*. The following procedure has been followed, for each frame k of each recording.

1. Determine the velocity of the i musician's head COG at frame k, as $\mathbf{v}_i^{(k)}$.
2. Evaluate the unit vector $\mathbf{d}_{i,ear}^{(k)} = \frac{\mathbf{e}-\mathbf{p}_i^{(k)}}{\|\mathbf{e}-\mathbf{p}_i^{(k)}\|}$ connecting the i musician's head COG to the *ear*, and determine the *radial* component $v_{\text{i,rad}}^{(k)} = \mathbf{v}_i^{(k)} \cdot \mathbf{d}_{i,ear}^{(k)}$ of $\mathbf{v}_i^{(k)}$, that is, the one along the direction $\mathbf{d}_{i,ear}^{(k)}$.
3. Compute the following quantity:
$S_{\text{rad}}^{(k)} = \sum_{i=1}^{4} \text{sign}(v_{\text{i,rad}}^{(k)})$, where $\text{sign}(v_{\text{i,rad}}^{(k)}) = \pm 1$;
This is a synchronization index that computes whether the movement changes of each musician (forward and backward to the SQ *ear*) happen simultaneously.

The second behavioral feature F_2 is defined as the percentage of frames where all the musicians move in a breathing coordinated manner, which is when $S_{\text{rad}}^{(k)} = \pm 4$:

$$F_2 = \%\text{ of frames for which } S_{\text{rad}}^{(k)} = \pm 4. \quad (2)$$

Again, this is computed for each single recording.

Finally, compared to [5, Subsection III.B], we have used a different definition of the feature F_2, because the one introduced in the present work is more suitable to measure the synchronization of the movements of the heads of the musicians (the old one tended to give too much importance to large values of the speed).

3.3 Convergence of a subset of 3 heads' directions toward the remaining musician

The third behavioral feature $\mathbf{F}_3$ is a vector made up of four components, one for each musician. It evaluates how the heads' directions of each subset of three musicians converge toward the head of the remaining one. The following procedure has been used, for each frame k of each recording,

1. For each musician $l = 1, \ldots, 4$, denote by $I_l^{(k)}$ the subset of the pairs (i, j) of musicians $(i, j = 1, \ldots, 4, i < j, i, j \neq l)$, different from l, for which the pairwise intersections defined in Sect. 3.1 exist at frame k, and by $|I_l^{(k)}|$ its cardinality.

If $I_l^{(k)}$ is nonempty, the position vector of the *point of partial convergence* (PoPC$_l$) associated with the musician l (see Fig. 2) is defined as

$$\mathbf{p}_{PoPC_l}^{(k)} = \frac{\sum_{(i,j)\in I_l^{(k)}} \mathbf{p}_{i,j}^{(k)}}{|I_l^{(k)}|}.$$

If $I_l^{(k)}$ is empty, the PoPC$_l$ is not defined at frame k. Interestingly, it follows by the definitions that the point of total convergence PoTC is the center of gravity of the set of all points of partial convergence PoPC$_l$, for the frames in which they are all defined.

2. Consider the distance $\|\mathbf{p}_{PoPC_l}^{(k)} - \mathbf{p}_l^{(k)}\|$ between the PoPC$_l$ associated with the musician l and his COG. When the PoPC$_l$ is not defined, the distance is set equal to its maximum value achieved in the frames of the recording for which the PoPC$_l$ exists.

Each component $F_{3,l}$ of the third behavioral feature $\mathbf{F}_3$ is defined as the median distance (inside each recording) between the PoPC$_l$ and the COG of the musician l:

$$F_{3,l} = \text{median of } \|\mathbf{p}_{PoPC_l}^{(k)} - \mathbf{p}_l^{(k)}\|. \quad (3)$$

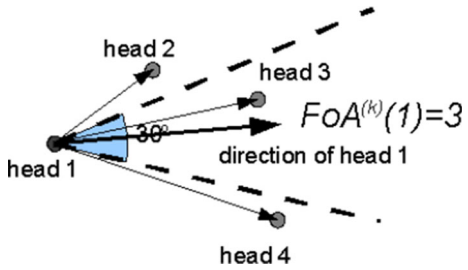


Fig. 4 Illustration of feature $\mathbf{F}_4$ measuring how much the head of each musician is directed toward each other musician. The figure shows an example of determination of the focus of attention (FoA) for the 1st violinist. In this case, the head of the violist (3rd player) is the one that minimizes the angle between the direction of the 1st violinist's head and any vector connecting the 1st violinist's head to any other head. Moreover, such minimum angle is less than the threshold, equal to 15°. The collection of FoAs is then used to compute the fourth feature $\mathbf{F}_4$ (see formula (4))

Here, similarly to the case of the PoTC considered in the feature F_1, the median distance has been considered instead of the mean, because it is less sensitive to outliers. This feature is also normalized dividing its components by the respective medians over all the recordings, in such a way that their ranges are comparable to the ones of the other features. Finally, the results of the data analysis in Sect. 4.3 refer to such a normalized feature.

3.4 Focus of attention (FoA) of single musician

The fourth set of behavioral features $\mathbf{F}_4$ is a matrix whose elements specify how much the head of each musician is directed toward each other musician. The following procedure has been followed, for each frame k of each recording and for each musician i (see Fig. 4 for an example).

1. Compute the angles between the direction of musician i's head and the vectors connecting his head to the each other's one, respectively.
2. If the minimum of these angles is smaller than a given threshold (set to the value 15 °, adapted from the literature [31]) and is achieved for the musician $\hat{j}$, then the head of musician i is considered as directed toward musician $\hat{j}$. Define $\hat{j}$ as the *focus of attention* of the musician i, $FoA^{(k)}(i) = \hat{j}$ (in the—unlikely—case of odds, one can choose at random one of the musicians that minimizes the angle in that frame, and define the associated number as the *focus of attention* of the musician i in that frame).
3. Otherwise, conclude that the head of the musician i is not directed toward any other musician in such frame. Define 0 as the *focus of attention* of i, $FoA^{(k)}(i) = 0$.
4. By definition, there are no frames for which $FoA^{(k)}(i) = i$ (no musician is directed toward himself).

Each element $F_{4,i,j}$ of the fourth behavioral feature $\mathbf{F}_4$ (which is defined for each single recording) is defined as the percentage of frames in which the *focus of attention* of the musician i is j (the possible values of i and j belong respectively to the set $\{1, 2, 3, 4\}$ and $\{0, 1, 2, 3, 4\}$):

$$F_{4,i,j} = \%\text{ of frames for which } FoA^{(k)}(i) = j. \quad (4)$$

4 Results of the data analysis on single features

This section describes the results obtained for each feature defined in Sect. 3, for both conditions A and B. All features were submitted to statistical tests to draw inferences on them. The distribution of data and their variances were first verified to select the most appropriate statistical tests. For all features, the obtained values did not follow a normal distribution according to the selected normality test (Kolmogorov-Smirnov). The variances were also not homogeneous according to the Levene statistical test. Specific non-parametric tests—which do not require the assumptions of a normal distribution and equal variances (of the residuals)—were used for the analysis of the various features. More precisely, Mann–Whitney U test was applied on the mean F_1, F_2, $\mathbf{F}_4$ feature values of all recordings taken together, for each condition A and B. A multi-level model (more precisely, a linear mixed model, LMM) was alternatively used for the $\mathbf{F}_3$ feature to consider an additional level in the analysis (subset of musicians) that could not be included otherwise. (possibly due to the small number of samples). Regarding the choice of the mean (with respect to all the recordings associated to the same performance condition) in some statistical tests performed in the current section, this is justified by the fact that such statistical tests evaluate whether two sets of samples come from populations with the same mean or not. In the following subsections, we report the cases in which the results were evaluated to be significant by the tests, and also the cases in which they were evaluated to be not significant (possibly due to the small number of samples).

4.1 Convergence of the heads' directions toward the ear

For the recordings considered in the data analysis, the parallelism condition mentioned in Sect. 3.1 actually never occurred,[6] due to the seat configuration of the players and the resolution of the instruments. Nevertheless, as in the definition of the feature F_1 we have considered intersections between half-lines, some of such intersections were not defined in some frames. This happened, for instance, when a pair of musicians looked at opposite directions. However, for each recording, the sets of intersections used to define the PoTC was not empty in each frame, so such a point was actually defined for all the frames, and the correction mentioned in Sect. 3.1 was not needed for the feature F_1 (a similar remark holds for all the PoPCs and the feature $\mathbf{F}_3$). Then, non-parametric Mann–Whitney U test showed that there was no significant difference for the mean values[7] of the first feature F_1 (U=5, p = 0.117) with respect to the recordings associated with each performance condition, see Fig. 5.

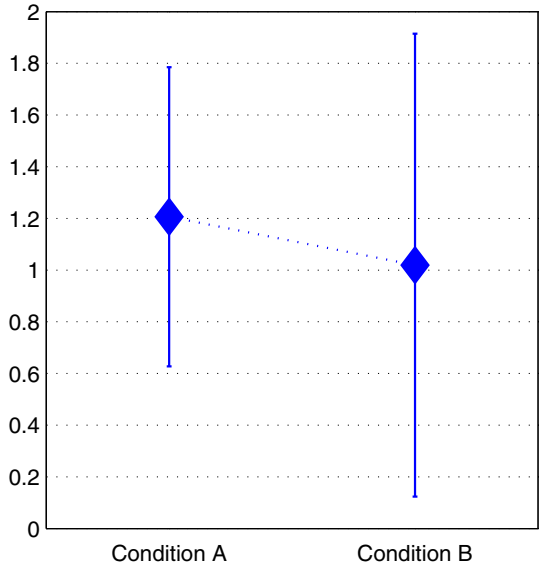


Fig. 5 Means and confidence intervals for the first feature F_1 (after normalization) in conditions A and B. Musicians' heads variability was higher in the *perturbed* condition (B) with respect to the *concert-like* one (A)

4.2 Joint movement dynamics of the heads toward the ear

For the second feature F_2, non-parametric Mann–Whitney U test showed that the difference between its mean values under conditions A and B was not significant ($U = 4$, $p = 0.972$). In both conditions, feature F_2 mean values and standard deviation were similar, revealing similar cohesion among musicians in their movements along the direction of the *ear* (Fig. 6).

[6] That condition was mentioned in Sect. 3.1 only to describe how the procedure could be modified in the unlikely case such a condition would occur.

[7] One can notice that there is no contradiction about the use of the median inside the definition of the feature F_1 in Sect. 3.1, and the use of the mean instead in the analysis described in this subsection. Indeed, the median among the frames of each recording was used in the definition of the feature F_1 in Sect. 3.1, whereas the mean in this subsection was computed at another level of the analysis, i.e., averaging the obtained values of the feature F_1 with respect to the recordings associated with the same performance condition. A similar remark holds for the feature $\mathbf{F}_2$.

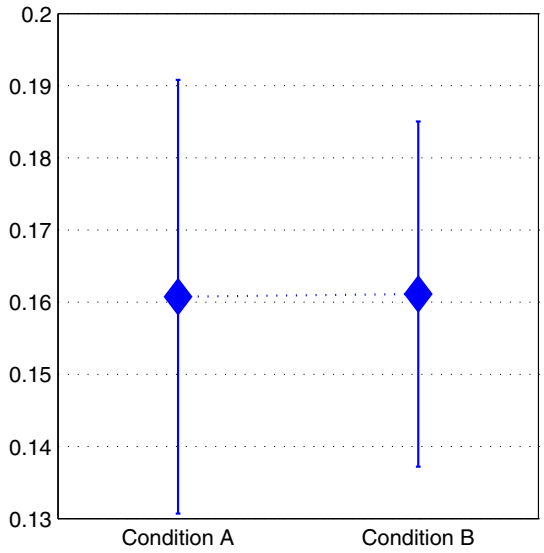


Fig. 6 Means and confidence intervals for the second feature F_2 in conditions A and B. One can notice similar values in both conditions showing that cohesion among musicians along the direction of the *ear* remain similar notwithstanding the perturbation

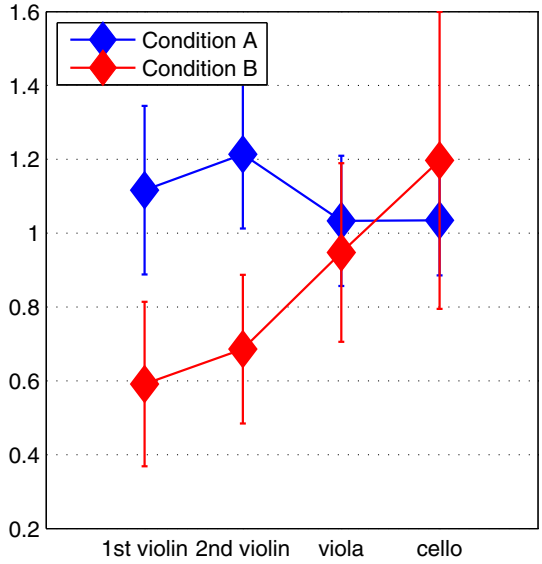


Fig. 7 Means and confidence intervals for the third feature $\mathbf{F}_3$ components (after normalization) in conditions A and B. One can notice that the distance between the 1st violinist and his associated point of partial convergence (PoPC_1) was significantly smaller in the *perturbed* condition (B) with respect to the *concert-like* one (A). A similar result was obtained for the 2nd violinist, whereas the opposite was observed for the cellist (who sits in front of the 1st violinist)

4.3 Convergence of a subset of three heads' directions toward the remaining musician

A LMM was chosen to compare musicians' third feature $\mathbf{F}_3$ values across conditions A and B to handle correlated data and unequal variance observed in the dataset. To control the inflation of type I error probability due to multiple comparisons, the Bonferroni correction was applied to adjust the α-value (the level of statistical significance). The LMM identified significant main effects of condition (A vs B), ($p < 0.001$). As shown in Fig. 7, the distance between the 1st violinist and his associated point of partial convergence (PoPC_1) decreased significantly from the *concert-like* condition (A) to the *perturbed* one (B), revealing how the 2nd violinist, violist and cellist's heads are converging toward him. As a side effect, the distance between the 2nd violinist and his associated point of partial convergence (PoPC_2) decreased significantly, whereas the distance between the cellist and his associated point of partial convergence (PoPC_4) increased significantly.

4.4 Focus of attention (FoA) of single musician

A statistical analysis was performed, investigating the values of the components of the fourth feature $\mathbf{F}_4$ related to the 1st violinist. These components quantify how much the 2nd violinist, the violist and the cellist are focusing on the 1st violinist in conditions A and B. A non-parametric Mann–Whitney U test revealed a significant difference in the mean values ($U = 8{,}282$, $p < 0.001$). Figure 8 shows pie charts summarizing the mean values of the components of the fourth feature $\mathbf{F}_4$ for all the musicians, in conditions A and B.

4.5 Questionnaire

Independent samples t tests were conducted to compare the ratings of *expressivity* and *cohesion*, in each performance condition A and B, as indicated by the four musicians themselves after each recording. Results (means and confidence intervals) are shown in Fig. 9. The difference in ratings was significant, $t(38) = 12.13$, $p < 0.001$ for *expressivity*, not for *cohesion* ($p = 0.07$). Interestingly, this rating of *cohesion* is consistent with the findings obtained in Sect. 4.2 for feature F_2, according to which cohesion was high in both conditions, despite the perturbation specific of condition B.

5 Classification results using combinations of features

As shown in Sect. 4, no single feature was sufficient to distinguish completely between conditions A and B. Therefore, in this section we explore the combinations of the features. More precisely, in the following we describe the results we obtained by applying machine learning techniques to automatically classify the data into the two classes (conditions) A and B.

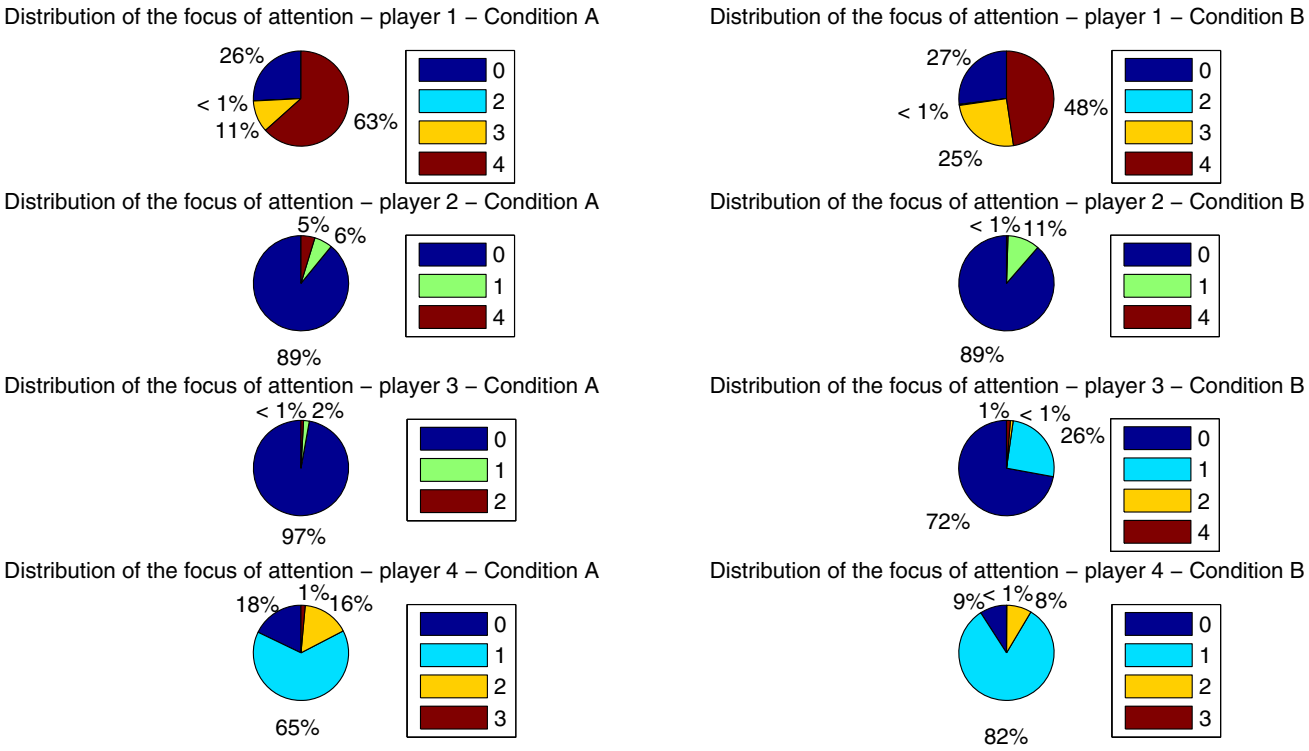


Fig. 8 Pie-chart representation of the mean values of the components of the fourth feature $\mathbf{F}_4$, in conditions A and B. One can notice that the 2nd violinist, violist and cellist focused their attention on the 1st violinist more in the *perturbed* condition (B) with respect to the *concert-like* one (A)

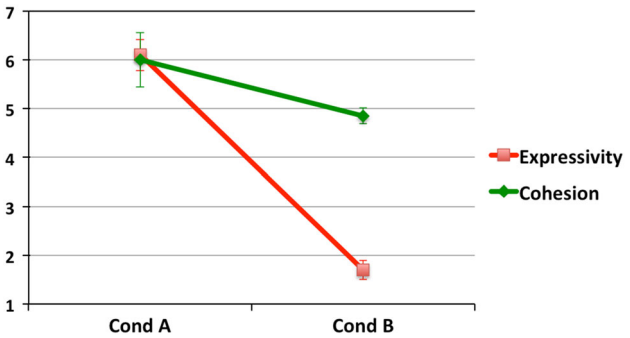


Fig. 9 Means and confidence intervals for the *expressivity* and *cohesion* items in the questionnaire, in conditions A and B. One can notice that *expressivity* was larger in the *concert-like* condition (A) with respect to the *perturbed* one (B), whereas *cohesion* was similar in the two cases

5.1 Data fragmentation

In the data analysis performed in this section, more punctual features were used. They were obtained fragmenting the recordings into smaller time windows, in order to get more dynamic data, together with a sufficient amount of training data. More precisely, two data analysis were performed, one with 4-s windows and 2-s interonset intervals,[8] and one with 2-s windows and 1-s interonset intervals. The two sizes of the time window were selected as they refer to two levels of segmentation of the music score, corresponding, respectively, to riff (2s) and music semiphrase (4s) levels of analysis. The 4 resulting features F_1, F_2, $\mathbf{F}_3$ and $\mathbf{F}_4$ (evaluated on each time window, adapting to it the procedure detailed in Sect. 3) include 22 sub-features (1 for F_1, 1 for F_2, 4 for $\mathbf{F}_3$, and 16 for $\mathbf{F}_4$), which were normalized. A first analysis showed that there was no clearly discriminating (sub)feature, as the error intervals for condition A and condition B intersected for all cases, nevertheless one feature resulted immediately to be useless: $F_{4,2,3}$, which was always 0 since in the database of recordings considered in the present work, the 2nd violinist never glanced in the violist's direction. This was expected due to player arrangement reasons, combined with the strong leader role assumed by the 1st violinist.

5.2 F-scores

A fivefold cross-validation technique was used in the following data analysis. More precisely, in the kth fold ($k = 1, \ldots, 5$), the test set was made up of the time windows cor-

[8] The interonset interval (IoI) is the lapse of time between the beginnings of two consecutive time windows. Since in this work it was chosen to be smaller than the length of the time windows, consecutive time windows always overlapped. Although this introduced an additional correlation between the features computed on different time windows, this was limited to consecutive time windows.

Table 2 Means, standard deviations and medians (over the folds) of the F-scores of the (sub)features for the binary classification problem between classes (conditions) A and B

Feat.	2s w. mean	2s w. std.	2s w. mdn.	4s w. mean	4s w. std.	4s w. mdn.
F_1	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$
F_2	0.01	$\simeq 0$	0.01	0.01	$\simeq 0$	0.01
$F_{3,1}$	71.56	80.56	34.34	67.15	56.51	42.72
$F_{3,2}$	108.20	26.83	116.45	83.17	25.11	79.27
$F_{3,3}$	9.68	4.84	10.72	7.60	2.68	8.65
$F_{3,4}$	25.43	46.80	2.04	24.64	47.41	0.94
$F_{4,1,0}$	0.52	0.47	0.37	0.35	0.33	0.20
$F_{4,1,2}$	2.50	0.71	2.65	2.04	0.49	2.13
$F_{4,1,3}$	49.90	15.93	53.16	35.69	11.77	37.17
$F_{4,1,4}$	34.06	11.63	38.41	24.21	8.75	27.21
$F_{4,2,0}$	0.49	0.68	0.14	0.30	0.41	0.10
$F_{4,2,1}$	6.66	5.95	4.51	3.87	3.54	2.60
$F_{4,2,3}$	NaN	NaN	NaN	NaN	NaN	NaN
$F_{4,2,4}$	18.54	4.52	17.90	11.79	2.64	11.05
$F_{4,3,0}$	106.98	35.70	121.67	59.87	19.93	68.74
$F_{4,3,1}$	112.02	22.96	122.05	63.47	12.56	69.41
$F_{4,3,2}$	1.98	2.51	0.03	1.22	1.57	0.02
$F_{4,3,4}$	5.52	1.84	6.91	3.49	1.15	4.37
$F_{4,4,0}$	19.24	8.23	17.87	13.44	5.95	11.96
$F_{4,4,1}$	44.51	16.08	52.22	27.76	10.40	32.90
$F_{4,4,2}$	13.56	9.03	10.10	8.31	5.73	6.01
$F_{4,4,3}$	5.62	1.73	6.03	3.75	1.29	3.92

The F-scores of the (sub)features depended on the durations of the time windows, but they were nearly in the same order of effectiveness for both cases

responding to the kth recording made under condition A and the kth recording made under condition B, whereas the training set was made up of all the time windows coming from the other 8 recordings. In such a way, for each fold, there was no overlap between training and test windows, since they were selected from different recordings. Then, for each fold, in order to assess the discrimination power of the features, their F-scores (Fisher scores) were computed. Given the training vectors x_k, $k = 1, \ldots, m$, the numbers of instances in class A and B (i.e., instances associated with each of the two classes) are denoted by n_A and n_B, respectively, so $m = n_A + n_B$ (such a number depends on the fold). Then, for each fold, the F-score of the ith feature is defined as follows [37]:

$$F(i) = \frac{(\bar{x_i}^A - \bar{x_i})^2 + (\bar{x_i}^B - \bar{x_i})^2}{\frac{1}{n_A - 1}\sum_{k=1}^{n_A}(x_{k,i}^A - \bar{x_i}^A)^2 + \frac{1}{n_B - 1}\sum_{k=1}^{n_B}(x_{k,i}^B - \bar{x_i}^B)^2},$$

where $\bar{x_i}$, $\bar{x_i}^A$, $\bar{x_i}^B$ are the average of the ith feature on the whole training set of the fold, and on its subsets corresponding to class A and class B, respectively; $x_{k,i}^A$ is the ith feature of the kth instance of class A in the training set of the fold, and $x_{k,i}^B$ is the ith feature of the kth instance of class B in the training set of the fold. The numerator evaluates the discrimination between the class A and class B sets, and the denominator evaluates the variability within the two sets. The F-score is easy to compute and generally effective in evaluating discriminating features.

The F-scores of the four main features ranked them in the following order: $\mathbf{F}_4$ (mostly, the violist's and the 1st violinist's foci of attention) and $\mathbf{F}_3$ (in particular, the components concerning the 2nd and the 1st violinists), then F_2 and F_1. The details are given in Table 2. The F-scores were computed for 2-s windows and 4-s windows, and some differences appeared, which was expected, as the features vary differently in time, and thus have different stabilities with respect to time. For example, the point of total convergence is expected to be more stable than the focus of attention for each single musician. Some features resulted to be more suitable to short-time analysis, while others resulted to be more effective for longer time windows. Finally, note that $F_{4,2,3}$, which was always 0, was useless for the discrimination between the two classes.

5.3 Support vector machine classification

Doing a classification test is a way to show that a model trained with the aim of discriminating between different classes has a good generalization capability on examples that were not used in the training phase. Those classification tests (which were performed 5 times, using the test sets of the dif-

ferent folds) are also called validation tests. Their results are described in the Sect. 5.4. In this subsection, we describe the procedure we followed to train the classifiers for each fold.

In order to classify the data into the two classes (conditions) A and B, a *SVM* was trained, using the parameter optimization strategies explained in [38].

The algorithm is as follows:

1. for each fold (generated according to procedure described in Sect. 5.2), do a parameter optimization (γ and C parameters of the *SVM*) with a polynomial kernel and a fivefold cross-validation on the X_{train} data, and train an optimized *SVM* classifier;
2. test the classifier on the set X_{test} of such fold;
3. repeat five times the procedure above to obtain an average test error with respect to the folds.

The last step means that a nested cross-validation is performed. Moreover, we also investigated the classification performances obtained using less features. In particular, the best 15, 11 and 9 features have been considered, according to the ranking of their F-scores in each fold (roughly, such choices correspond to F-score thresholds equal to 1, 5, and 10, respectively). Also in this case, for each fold, the γ and C parameters of the *SVM* were optimized through a fivefold cross-validation on the X_{train} data of the fold.

5.4 Classification results

Table 3 summarizes the results of the classification tests that were obtained for the procedure described in Sect. 5.3. For such tests, means and standard deviations with respect to the folds were computed. A great percentage of time windows was correctly classified, always with performance better than chance (50 % for the balanced data considered in this paper), even when using only nine features. The algorithm showed to be most effective when all the 22 features were used, but the results were in a similar scale when using only the 15 best features selected according to their F-score ranking. Finally, the results worsened as the set of features used for classification was further reduced. When used together, the 22 features we have defined in the paper demonstrated therefore to be efficient for classification, and led to a percentage of about 90 % successfully classified test windows. One can also notice that the classification results obtained using 2- and 4-s windows were quite similar.

Table 3 Classification results (in percentages of correct classification on the test set, averaged on fivefolds) for the two sizes of the time windows and different numbers of selected features (according to the F-score in each fold)

# Feat.	2s wind. classif. (%)	4s wind. classif. (%)
22	89.9 ± 2.4	90.3 ± 3.4
15	87.0 ± 3.8	87.6 ± 6.1
11	72.6 ± 4.4	74.4 ± 6.9
9	63.1 ± 4.9	65.9 ± 6.1

Also the standard deviation with respect to the folds is shown

6 Conclusions

Playing music with others represents one of the most engaging and expressive experience [1]. The findings of this paper show that a set of behavioral features could be implemented to automatically distinguish between a highly satisfying engaging and expressive type of performance versus a less satisfying expressive performance. Specifically, in the global analysis presented in Sect. 4, we found that: features F_1 and F_2 revealed that a cohesive performance in both performing conditions A and B was obtained (see Sects. 4.1, 4.2); feature $\mathbf{F}_3$ showed that in *perturbed* condition B, one point of interest plays a central role: the point of partial convergence PoPC_1 associated with the 1st violinist (see Sect. 3.3); features $\mathbf{F}_3$ and $\mathbf{F}_4$ enabled to distinguish between *concert-like* condition A from *perturbed* one B, the results obtained from feature analysis were consistent with the results of the questionnaire (see Sect. 4.5). Finally, we also combined the features to perform binary classification using an *SVM* classifier (see Sect. 5), also making a ranking of the features in terms of their F-scores. The dynamism of the performance was well rendered for both 2- and 4-s windows, and the classification algorithm showed a very high success rate (see Sects. 5.1, 5.2, 5.3, 5.4). One of the motivations of the local analysis made in Sect. 5 was the larger number of data used in such an analysis, with respect to the global one performed in Sect. 4. This was due to the data fragmentation in time windows. Interestingly, the features whose discriminative power was evaluated as large (respectively, small) in the global analysis performed in Sect. 4, were also evaluated in a similar way in the local analysis performed in Sect. 5.2. Finally, an *SVM* was trained to perform the automatic discrimination of the time windows extracted from the various performances. The experiments showed that the selected 22 local features were able to separate well the data obtained on the time windows associated with the performances made under conditions A and B. The experiments also showed that using less features produced often a less accurate classification, but even a subset of 15 features produced similar results than the whole set of features.

The results of this paper confirm in a quantitative way previous studies on music ensemble performance (e.g., [17]). It has actually been suggested that musicians pay attention to other performers' heads to better predict their upcoming actions. This is particularly obvious when the behavior of a

musician is difficult to predict. Indeed, condition *B* of the present study sets the 1st violinist in a privileged position, providing him with information that could be transmitted to the other musicians mainly through his movement, and constraining all the other musicians to follow him tightly to maintain the group cohesion. More generally, the present study highlights the potential of the SQ scenario as a test case to study group behavior in social emotionally-engaged and creative activities in ecological settings.

Although an evaluation of the expressivity and cohesion by an audience was not performed in this work (whose main objective was to find suitable features able to discriminate between the two performance conditions, independently from how such conditions were perceived by the public), we mention that in other works on string quartets we took into account such a perception. For instance, in the experiment described in [39], the musicians were asked to play the same musical fragment in a solo condition and then with the rest of the quartet, and the public was asked to discriminate between such two conditions looking only at the MoCap data associated with one player. In such a case, the difference between the two conditions resulted to be not obvious to the untrained observers, while musician observers were able to distinguish between the two conditions.

The data analysis presented in this paper was performed at two different levels. In the first simpler level of analysis, the possible dependence of the features on time was not considered, and global features were considered. In the second level of analysis, the features were evaluated on shorter time windows (2 and 4 s long), which allowed to use dynamic models for the analysis, which were more refined, due to the increased number of data obtained by segmenting the original data. Of course, an even more sophisticated analysis may be performed, e.g., one may use some training data to fit the parameters of a time-series model to the data. In the previous work [21], for instance, such a method was used, combined with Granger's causality, to measure to what an extent a time series associated with one player (in that case, the time series of the distance of the head of that player from the "ear" of the quartet), influenced the corresponding time series associated with another player. As a possible extension of the present work, a similar analysis may be performed using the features considered in this paper. Other possible extensions in the analysis include: the detection of possible nonlinear dependencies among the individual features of different members of the SQ; the application of the methodology of the present work to other SQs and other settings (e.g., musicians in orchestras, or other small groups of highly skilled people, not necessarily musicians, such as dancers or athletes); the use of other tools, such as Google Glass, to obtain possibly better estimates of the focus of attention, and better measurements of the head movements (possibly using suitable image processing techniques).

Acknowledgments The project SIEMPRE acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open Grant Number: 250026-2.

References

1. Eerola T, Vuoskoski JK (2013) A review of music and emotion studies: approaches, emotion models, and stimuli. Music Percept Interdiscip J 30(3):307–340
2. Vinciarelli A, Pantic M, Heylen D, Pelachaud C, Poggi I, D'Errico F, Schroder M (2012) Bridging the gap between social animal and unsocial machine: a survey of social signal processing. IEEE TAC 3(1):69–87
3. Seddon F, Biasutti M (2009) A comparison of modes of communication between members of a string quartet and a jazz sextet. Psychol Music 37(4):395
4. Gilboa A, Tal-Shmotkin M (2010) String quartets as self-managed teams: an interdisciplinary perspective. Psychol Music 40(1):19–41
5. Glowinski D, Gnecco G, Camurri A, Piana S (2013) Expressive non-verbal interaction in string quartet. In: Proc. of the fifth IEEE conference on affective computing and intelligent interaction (IEEE ACII 2013), pp 233–238
6. Cadoz C, Wanderley MM (2000) Gesture: music, chapter trends in gestural control of music. Ircam/Centre Pompidou, Paris, pp 71–94
7. Chadefaux D, Wanderley M, Le Carrou J-L, Fabre B, Daudet L (2012) Experimental study of the musician/instrument interaction in the case of the concert harp. In: Proceedings of acoustics
8. Wanderley MM (2002) Quantitative analysis of non-obvious performer gestures. Proc Gest Workshop 2001(2):241–253
9. Davidson JW (1993) Visual perception of performance manner in the movements of solo musicians. Psychol Music 21:103–113
10. Davidson JW (1994) What type of information is conveyed in the body movements of solo musician performers? J Hum Mov Stud 6:279–301
11. Castellano G, Mortillaro M, Camurri A, Volpe G, Scherer K (2008) Automated analysis of body movement in emotionally expressive piano performances. Music Percept 26:103–120
12. Palmer C, Koopmans E, Carter C, Loehr JD, Wanderley M (2009) Synchronization of motion and timing in clarinet performance. In Proceedings of the Second International Symposium on Performance Science
13. Varni G, Volpe G, Camurri A (2010) A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. IEEE Trans Multimed 12:576–590
14. D'Ausilio A, Badino, L, Li Y, Tokay S, Craighero L, Canto R, Aloimonos Y, Fadiga L (2012) Leadership in orchestra emerges from the casual relationships of movement kinematics. PLoS one 7. doi:10.1371/journal.pone.0035757
15. Gnecco G, Badino L, Camurri A, D'Ausilio A, Fadiga L, Glowinski D, Sanguineti M, Varni G, Volpe G (2013) Towards automated analysis of joint music performance in the orchestra. In: Arts and technology, third international conference, ArtsIT 2013, Milan, Italy, March 21–23, 2013, revised selected papers. Lecture Notes of the institute for computer sciences, social informatics and telecommunications engineering (LNICST) series, vol 116. Springer, Berlin, pp 120–127
16. 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
17. Davidson JW, Good JMM (2002) Social and musical co-ordination between members of a string quartet: an exploratory study. Psychol Music 30(2):186

18. Glowinski D, Coletta P, Volpe G, Camurri A, Chiorri C, Schenone A (2010) Multi-scale entropy analysis of dominance in social creative activities. In: ACM Multimedia MM10. ACM, Firenze, pp 1035–1038
19. Keller PE, Appel M (2010) Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles. Music Percept 28(1):27–46
20. Poggi I (2006) Body and mind in the pianist's performance. In: Proceedings of the 9th international conference on music perception and cognition, pp. 1044–1051
21. Glowinski D, Badino L, Ausilio A, Camurri A, Fadiga L (2012) Analysis of leadership in a string quartet. In: Third international workshop on social behaviour in music at ACM ICMI 2012
22. D'Ausilio A, Badino L, Li Yi, Tokay S, Craighero L, Canto R, Aloimonos Y, Fadiga L (2011) Communication in orchestra playing as measured with granger causality. In: Proceedings of the INTETAIN conference (intelligent technologies for interactive entertainment), Genova, May 2011
23. Luck G, Toiviainen P (2006) Ensemble musicians' synchronization with conductors' gestures: an automated feature-extraction analysis. Music Percept 24(2):189–200
24. Poggi I (2011) Music and leadership: the Choir Conductor's multimodal communication. John Benjamins Pub Co, Amsterdam, pp 341–353
25. Glowinski D, Camurri A, Volpe G, Noera C, Cowie R, McMahon E, Knapp B, Jaimovich J (2008) Using induction and multimodal assessment to understand the role of emotion in musical performance. In: Proceedings of the 2008 conference on emotion in human–computer interaction. Liverpool John Moores University, Liverpool
26. Schubert E (2001) Continuous measurement of self-report emotional response to music. In: Juslin PN, Sloboda JA (eds) Music and emotion: theory and research. Series in affective science. Oxford University Press, New York, NY, pp 393–414
27. Glowinski D, Torres-Eliard K, Chiorri C, Camurri A, Grandjean D (2012) Can naive observers distinguish a violinist's solo from an ensemble performance? A pilot study. In: Third international workshop on social behaviour in music at ACM ICMI 2012
28. Glowinski D, Dael N, Camurri A, Volpe G, Mortillaro M, Scherer K (2011) Toward a minimal representation of affective gestures. IEEE Trans Affect Comput 2:106–118
29. Dahl S, Bevilacqua F, Bresin R, Clayton M, Leante L, Poggi I, Rasamimanana N (2009) Gestures in performance. Sound Mov Mean Music Gest
30. Gnecco G, Glowinski D, Camurri A, Sanguineti M (2013) On the detection of the level of attention in an orchestra through head movements. Int J Arts Technol (to appear)
31. Stiefelhagen R (2002) Tracking focus of attention in meetings. In: Proceedings of the fourth IEEE international conference on multimodal interfaces. IEEE, New York, pp 273–280
32. Stiefelhagen R, Zhu J (2002) Head orientation and gaze direction in meetings. In: CHI '02 extended abstracts on human factors in computing systems, CHI EA '02. ACM, New York, pp 858–859
33. Stiefelhagen R, Yang J, Waibel A (2002) Modeling focus of attention for meeting indexing based on multiple cues. IEEE Trans Neural Netw 13(4):928–938
34. Camurri A, Dardard F, Ghisio S, Glowinski D, Gnecco G, Sanguineti M (2014) Exploiting the Shapley value in the estimation of the position of a point of interest for a group of individuals. Procedia Soc Behav Sci 108:249–259
35. Ba S, Odobez J-M (2006) A study on visual focus of attention recognition from head pose in a meeting room. In: Renals St, Bengio S, Fiscus JG (ed) Machine learning for multimodal interaction. Lecture notes in computer science, vol 4299. Springer, Berlin, pp 75–87
36. Hung H, Gatica-Perez D (2010) Estimating cohesion in small groups using audio–visual nonverbal behavior. IEEE Trans Multimed 12(6):563–575
37. Chang Y-W, Lin C-J (2008) Feature ranking using linear SVM. In: JMLR workshop and conference proceedings: causation and prediction challenge at WCCI 2008, vol 3, pp 53–64
38. Chen Y-W, Lin C-J (2006) Combining SVMs with various feature selection strategies. Feature extraction. In: Studies in fuzziness and soft computing, vol 207. Springer, Berlin, pp 315–324
39. Glowinski D, Mancini M, Cowie D, Camurri A (2013) How action adapts to social context: the movements of musicians in solo and ensemble conditions. In: Proc. of the fifth IEEE conference on affective computing and intelligent interaction (IEEE ACII 2013), pp 294–299