ANALYSIS OF SOCIAL INTERACTION IN MUSIC PERFORMANCE WITH SCORE-INDEPENDENT AUDIO FEATURES

Gualtiero Volpe, Giovanna Varni, Barbara Mazzarino, Silvia Pisano, Antonio Camurri
InfoMus - DIST - University of Genova
gualtiero.volpe@unige.it

ABSTRACT

Research on analysis of expressive music performance is recently moving its focus from a single player to small music ensembles, extending the analysis to the social interaction among the members of the ensemble. A step in this direction is the definition and the validation of a set of score-independent audio features that enable to characterize the social interaction in the ensemble, based on the analysis of the music performance. This paper focuses on the analysis of four different performances of a same music piece performed by a string quartet. The performances differ with respect to factors affecting the social interaction within the ensemble. The analysis aims at evaluating whether and to what extent a set of consolidated score-independent audio features, already employed for analysis of expressive music content and particularly suitable for string instruments, enable to distinguish among such different performances.

1. INTRODUCTION

Automatic analysis of music performance focused for long time on a single player. The aim was usually to identify a collection of audio and music features characterizing music performance in order to carry out various kinds of analysis and classification (e.g., genre, mood classification, and so on). Features concern several aspects: from low-level descriptions of the audio signal (e.g., energy-related features, spectral features, and so on), features derived from auditory models, features concerning higher-level music structures (e.g., articulation, phrasing). A typical example is the classification of the expressive content conveyed by the same music piece performed with different expressive intentions. In this case, this expressive content is represented and classified using different approaches, e.g., categorical approaches where it is characterized by expressive labels (e.g., the basic emotions) and dimensional approaches exploiting multidimensional spaces (e.g., the consolidated valence-arousal space), see for example [1][2][3][4][5].

A recent research trend shifts the focus to the *social dimension of music*. For example, Keller and colleagues analyzed piano duets and found that pianists were better at synchronizing with their own than with others performances, and they were able to recognize their own recordings [6]. As referred by Trehub [7] music perception and performance are basic aspects of human being development, the study of which enables a better and a deeper understanding of the emotional development and group social interaction. Further, music group performances such as, for example, ensemble performance and choir performance involve coordination of movements and alignment of mental states, and require cooperation due to a shared goal, division of roles, and monitoring of progresses (e.g., [8][9]). In such a framework, automatic analysis of music performance needs to be improved with models and algorithms to probe and quantify the social interaction between the members of an ensemble or between the musicians and the audience.

The EU-ICT-FET Project SIEMPRE (2010-2013, see also: www.infomus.org/siempre) investigates non-verbal creative communication, in terms of entrainment, emotional contagion, co-creation and leadership, within groups of performers and audience. Some previous studies, forming the background of SIEMPRE and based on movement and gesture analysis, already explored the emergence of some of these phenomena [10][11]. However, a multimodal approach to music is needed and an important issue in this direction is to identify score-independent features that can describe ensemble performances characterized by differences in the social interaction of the members of the group.

This paper considers a set of score-independent audio features that already proved to be significant to distinguish between performances having different expressive and sensorial intentions [2]. The purpose of the work is to find out whether and to what extent the same features can also be used for analysis of social interaction in a small music ensemble, namely a string quartet. To this aim, the features were extracted and analyzed from four performances differing with respect to factors affecting the social interaction within the ensemble, with a particular focus on the functional roles of the players.

Section 2 briefly introduces the audio features that were taken into account. Section 3 describes the experiment that provided the data set for testing the features and the obtained results.

2. AUDIO FEATURES

Research on social interaction in music performance needs, as a first step, to identify a music ensemble which is suitable as a test-bed for investigation. For example, preliminary studies on violin duo performances provided encouraging results with respect to the possibility of developing techniques for the analysis of two relevant social signals: the level of synchronization established among the behaviors of each single member of a group and the emergence of functional roles (e.g., a leader) [12][10]. Analysis of the duo performance, however, suggested that synchronization and, especially, leadership may be better assessed with larger groups of players. In this direction, a quartet seemed an ideal ensemble, big enough to clearly display the phenomena under investigation, but not so big to become too complex for experimental set-up and data analysis. The focus of this work is thus on string quartets and required the identification of audio features particularly suited for string instruments.

Social interaction within a music ensemble (e.g., the emergence of a leader) may either follow the indications provided in a score, or may be the result of the application of specific performance techniques, or may arise from the internal organization of the ensemble, refined and tuned in many sessions and rehearsals where musicians play together. Social signals also emerge in performances that are not characterized by an exactly predefined score, such as improvisation. Score-independent audio features would allow to analyze social interaction within the ensemble without the need of the knowledge of the score.

Mion and De Poli [2] tested several score-independent audio features for their effectiveness in classifing performances which differ with respect to expressive and sensorial intentions. They asked three professional performers of violin, flute, and guitar to play in order to convey different expressive intentions, described by affective (happy, sad, angry, and calm) and sensorial (light, heavy, soft, and hard) adjectives. Using a sequential feature selection procedure followed by a Principal Component Analysis they identified 5 features yielding a high percentage of correct classification for the violin performances. These include both local features, computed on sliding time windows, and event features, computed on single events, segmented from the audio stream and identified by their onsets and their offsets.

Local features include:

- *Roughness* (R), or Sensory Dissonance, a feature characterizing the texture of a sound in terms of impure or unpleasant qualities. Such a sensation is associated with the physical presence of beating frequencies in the auditory stimulus. Leman and colleagues [13] developed a technique, based on auditory modeling, for computing roughness in terms of the energy provided by the neural synchronization to beating frequencies. As such, roughness is computed by applying a Synchronization Index Model to the output of an auditory peripheral model. The IPEM Toolbox for auditory-based musical analysis [14] provides a Matlab implementation of roughness on top of the auditory periferal model of Van Immerseel and Martens [15].

- *Residual Energy* (RE) describes the stochastic residual of the audio signal, obtained by removing the deterministic sinusoidal components [16]. Residual energy can be computed over different frequency regions, however, as Mion and De Poli show [2], the residual energy in the frequency range above 1805 Hz (RE*h*) is particularly suited for the analysis of string instruments. RE*h* is thus computed as the residual energy ratio in such a frequency band:

$$REh = \frac{\sum_{j \in H} |X_R(j)|^2}{\sum_{k=1}^{N/2-1} |X_R(k)|^2}. \tag{1}$$

  where H is the set of spectrum bins corresponding to frquencies higher than 1805 Hz and X_R is the spectrum of the residual component of the signal.

Event features are computed on single events in the audio stream. Onset detection is performed by using the algorithm proposed in [17]. Offsets are detected as suggested in [2] when the root-mean-square (RMS) of the temporal envelop of the audio signal falls by the 60% from its previous maximum value. Event features include:

- *Notes per second* (NPS), computed by dividing the number of onsets by the duration of the analysis window. In [2] analysis is performed with windows of 4-s duration and 3.5-s overlap, so that the window size allows to include a reasonable number of events, and it corresponds roughly to the size of the echoic memory.

- *Attack time* (A), computed as the time required to reach the RMS peak, starting from the onset instant.

- *Peak sound level* (PSL) computed as the maximum value of the RMS within the event, i.e., $PSL = max(RMS(t))$.

3. ANALYZING SCORE-INDEPENDENT AUDIO FEATURES FROM A STRING QUARTET

3.1 Design and Material

The above-mentioned audio features were here applyed to analyze the social interaction among players. The features were extracted from the multi-track recordings of a professional string quartet, Quartetto di Cremona. The recordings were carried out at Casa Paganini - InfoMus, in occasion of a concert of Quartetto di Cremona at the Opera House concert season, and they were performed in an environment very similar to a concert hall, with technology and scientists participating in the studies hidden in the upper level room. In such a way, an effective ecological environment with no perturbation on the investigated phenomena was set-up.

Players were asked to perform the first movement (*Allegro*) of the Streichquartet No. 14 by Schubert in four different conditions:

- *Regular condition*: players play as in a regular concert performance;

- *Switch condition*: the first violin plays the score of the second violin and viceversa;

- *Functional condition*: players are asked to follow a metronome and focus only on the gestures that are directly needed in the sound production process;
- *Over-expressive condition*: players emphasize gestures and affective intentions.

These four conditions were selected in order to emphasize variations in the social interaction. The regular condition is the reference condition, where the quartet plays as in a regular concert, thus applying all the usual mechanisms and techniques they learned and tuned. The switch condition introduces an explicit external action (the switch of the score of the first and second violin) to affect the normal social relationships between the members of the ensemble (e.g., a possible change of leadership). The functional (metronomic) condition operates on the social interaction in two different ways: on the one hand, it imposes a kind of external leader, the metronome; on the other hand it inhibits the expressive content conveyed by the piece. The over-expressive condition exaggerates the expressive content the music conveys, thus possibly requiring a higher degree of cohesion and synchronization in the group. Note that, even if the conditions also differ with respect to expressive content (in particular the functional and the over-expressive conditions), here the focus is not on the expressive content, but rather on social interaction. That is, the variation in the amount of expressive content to be conveyed is exploited as a way to affect the established social mechanisms of the ensemble in order to make the social variables emerge more evidently. Nevertheless, the presence of such variations of expressive content is a further motivation for choosing score-independent audio features that already prooved to be significant in distinguishing between different expressive performances.

Each condition was repeated two times - two performances - with a short break in between them. Table 1 shows the experimental protocol. This resulted in the same piece repeated 8 times: 2 regular performances, 2 switch performances, 2 functional performances, and 2 over-expressive performances.

No.	Condition	Performance
1	Regular	I
2	Regular	II
3	Switch	I
4	Switch	II
5	Functional	I
6	Functional	II
7	Over-expressive	I
8	Over-expressive	II

Table 1: The order conditions were performed.

The recordings used in this work are part of a wider study aiming at measuring and evaluating social features in music ensemble performances, with particular reference to synchronization and leadership. Such a wider study also includes measures from physiological sensors and visual recordings. Initial results are discussed in [10].

3.2 Analysis and Results

The score was divided into 5 Parts based on salient points such as, for example, pauses, attacks, and changes in the dynamics. For each single Part and for each audio track of each player, the selected score-independent audio features were extracted using the same settings (e.g., window size, hop size, and so on) as indicated in [2]. These features were the dependent variables of the experiment, whereas the independent variables were *Condition* and *Performance*. The mean value of each feature in each Part was chosen as a synthetic descriptor for the Part. The effect of *Condition* and *Performance* and their combined effect were assessed with an RM two-way within subjects ANOVA on each of the features. The analysis was carried out on each of the five segmented Parts: 25 RM two-way within subjects ANOVA (5 audio features x 5 Parts).

Condition had no significant effect on any feature, whereas *Performance* had effect on RE Part III ($F = 9.7, p < 0.05$) and Part IV ($F = 25.94, p < 0.05$), NPS Part I ($F = 4.16, p < 0.05$), PSL Part II ($F = 9.24, p < 0.01$), and A Part II ($F = 4.72, p < 0.05$). A significant interaction *Condition* and *Performance* was found on NPS Part I ($F = 4.16, p < 0.05$), PSL Part II ($F = 6.13, p < 0.05$), A Part I ($F = 3.94, p < 0.01$), and Part II ($F = 4.19, p < 0.05$). Both the main factors and their interaction do not seem to affect the *R* feature. A further Bonferroni corrected post-hoc analyses could assess specific differences among the effects. Table 2 summarizes the number of Parts in which each feature was statistically significant with respect to the main factors and their interactions.

Feature	Factor(s)	No. Parts
NPS	Condition	0
NPS	Perfomance	1
NPS	Condition*Performance	1
R	Condition	0
R	Performance	0
R	Condition*Performance	0
A	Condition	0
A	Performance	1
A	Condition*Perfomance	2
RE	Condition	0
RE	Performance	2
RE	Condition*Performance	0
PSL	Condition	0
PSL	Performance	1
PSL	Condition*Performance	1

Table 2: No. of Parts in which each feature was significant.

Figure 1 depicts the interaction plot between *Condition* and *Performance* for the NPS feature in Part I (upper panel), and for the PSL feature in Part II (lower panel), respectively. In both the panels the means change over *Condition* in different ways, resulting in lines having a different slope. NPS increases and decreases across *Condition*, whereas PSL slowly increases. This change of slope reveals that interaction between the variables is significant as confirmed by the previous numeric analysis.

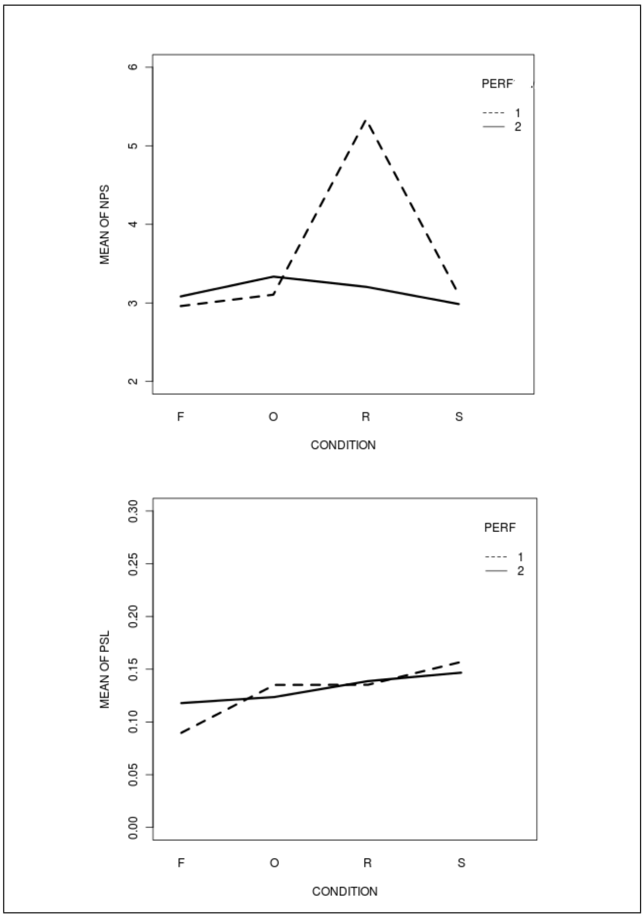

Figure 1: Upper panel: the interaction between *Condition* and *Performance* for NPS (Part I). Lower panel: the interaction between *Condition* and *Performance* for PSL (Part II). x-axis shows the *Condition* in alphabetic order: F: functional, O: over-expressive, R: regular, S: switch of score.

The analysis does not show significant between-subjects differences for each feature and for each Part. The interaction plot of Figure 2 exemplifies this in the case of feature A Part II. The variables *Instrument* and *Performance* are not significant: the lines are rather close together and there is no change over *Performance* (upper panel). In the lower panel the lines are also close together and they are parallel except for the line labeled as vio4 (vio4 is the cello player).

In conclusion, the selected features do not seem to provide enough information to distinguish among conditions. A motivation for this may be that the experiment involved a professional quartet, who is able to promptly react to possible changes and perturbations in the social interaction, so that the audio result does not fully reproduce such changes. In order to assess this hypothesis, experiments should be also performed with non-professional music players (e.g., a student quartet). Indeed, in a previous work on the same recordings, a significant effect was found on beat [10]. Results show that in some cases *Performance* is significant, i.e., there is a significant change in the audio features between the first and the second performance of the same condition. This may be due to a lower stress in the players following the acquired knowledge of the specific condition (the quartet was not aware of each single condition before the experiment). However, such an effect still needs to be further investigated.

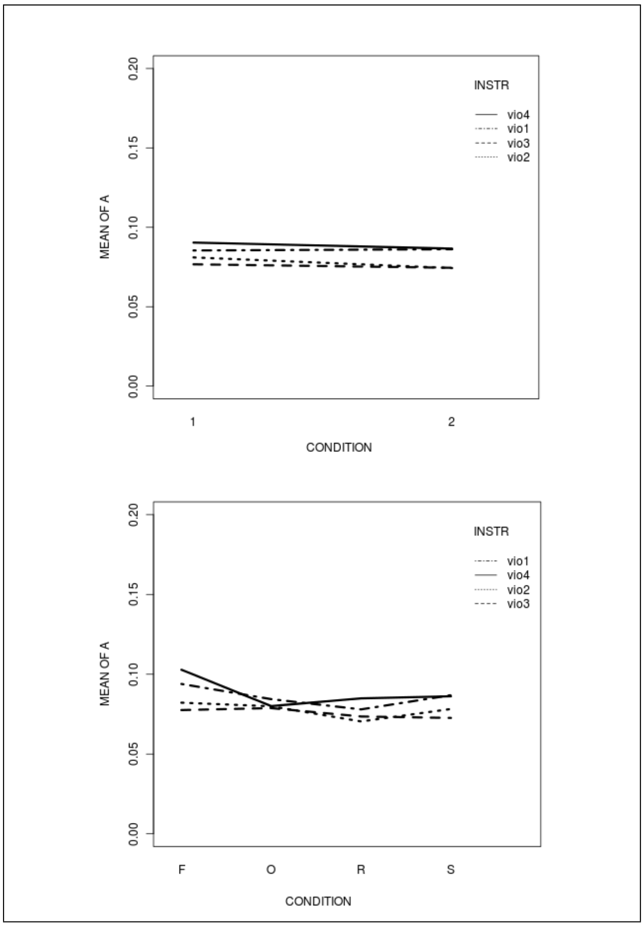

Figure 2: Upper panel: the interaction between player and performance for A; the x-axis shows the two performances. Lower panel: the interaction between players and condition for A; the x-axis shows the four conditions in alphabetic order: F: functional, O: over-expressive, R: regular, S: switch of score. The legend shows the four players: vio1 is the first violin, vio2 is the second violin, vio3 is the viola, and vio4 is the cello.

4. CONCLUSIONS

This paper addressed the identification and the analysis of score-independent audio features able to catch the social dimension of music. The selected features already proved able to distinguish performances acted with different expressive intentions. However, the carried out analysis revealed that such a set of features cannot provide sufficient information to distinguish among the different social conditions tested in this experiment. This result may either depend on the conditions, i.e., the introduced perturbations may have been too weak to produce an appreciable change in the music played by the quartet or on the features that may be unable to capture the possible variations induced by the perturbations.

Nevertheless, the output of the work still suggests some possible perspectives for future research. One direction concerns the experiments to be performed and the conditions to be tested. A comparative analysis of professional and non-professional quartets could be carried out in order to better identify the perturbations that are likely to have a major impact on the social interaction within the ensemble. In this framework, it may be useful to have more quartets and to also use questionnaires for better assessing to what extent musical skills are related with and affect social interaction. Further possibilities include, for example, comparing musicians that usually play together and musician that do not, testing a condition where only one player is instructed to change her way to play during the performance, using less familiar music pieces making the performance more similar to improvisation.

Another direction is related to the features to be used for analysis. A deeper investigation on the possible dependence of the features from the music instrument is needed. If features are independent from the instruments, as the preliminary results obtained here seem to suggest, these may be used to improve e.g., analysis of leadership, making it less dependent on the music instrument. Moreover, as the previous work on beat analysis shows [10], the set of features can be changed/extended by including new ones that explicitly take into account the temporal dynamics and rhythmic aspects of music. Multimodal integration with motion capture data is also being investigated.

Some of the issues above are currently addressed in a study involving a string quartet of students of the Music Conservatory of Genova. New experiments are planned within the SIEMPRE Project in the near future.

Acknowledgments

This work has been partially supported by the SIEMPRE Project. 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. The authors thank Quartetto di Cremona and the colleagues at Casa Paganini - InfoMus.

5. REFERENCES

[1] G. Widmer and W. Goebl, “Computational models of expressive music performance: The state of the art,” *Journal of New Music Research*, vol. 33(3), 2004.

[2] L. Mion and G. D. Poli, “Score-independent audio features for description of music expression,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16(2), 2008.

[3] A. Camurri, G. D. Poli, A. Friberg, M. Leman, and G. Volpe, “The mega project: analysis and synthesis of multisensory expressive gesture in performing art applications,” *Journal of New Music Research*, vol. 34(1), 2005.

[4] W. Goebl, S. Dixon, G. D. Poli, A. Friberg, R. Bresin, and G. Widmer, “Sense in expressive music performance: Data acquisition, computational studies, and models,” in *Sound to Sense, Sense to Sound - A State of the Art in Sound and Music Computing*, 2007.

[5] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. R. Scherer, “Automated analysis of body movement in emotionally expressive piano performances,” *Music Perception*, vol. 26(2), 2008.

[6] P. E. Keller, G. Knoblich, and B. H. Repp, “Pianists duet better when they play with themselves: On the possible role of action simulation in synchronization,” *Consciousness and Cognition*, vol. 16, 2007.

[7] S. Trehub, “The developmental origins of musicality,” *Nature Neuroscience*, vol. 6(7), 2003.

[8] M. Tomasello and M. Carpenter, “Shared intentionality,” *Developmental Science*, vol. 10, 2007.

[9] S. Koelsh, “Towards a neuronal basis of music-evoked emotions,” *Trends Cogn Sci*, vol. 14(3), 2010.

[10] G. Varni, G. Volpe, and A. Camurri, “A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media,” *IEEE Transactions on Multimedia*, vol. 12(6), 2010.

[11] D. Glowinski, P. Coletta, C. Chiorri, A. Camurri, G. Volpe, and A. Schenone, “Multi-scale entropy analysis of dominance in social creative activities,” in *Proc. of the ACM Multimedia 2010 Conference (MM ’10)*, 2010.

[12] G. Varni, A. Camurri, P. Coletta, and G. Volpe, “Emotional entrainment in music performance,” in *Proc. 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG2008)*, 2008.

[13] M. Leman, “Visualization and calculation of roughness of acoustical musical signals using the synchronization index model (sim),” in *Proc. of the 2000 COST G-6 Conference on Digital Audio Effects (DAFX-00)*, 2000.

[14] M. Leman, M. Lesaffre, and K. Tanghe, “A toolbox for perception-based music analysis.” Institute for Psychoacoustics and Electronic Music (IPEM), 2005.

[15] L. V. Immerseel and J. Martens, “Pitch and voiced/unvoiced determination with an auditory model,” *Journal of Acoustical Society of America*, vol. 91(6), 1992.

[16] N. Laurenti and G. D. Poli, “A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15(2), 2007.

[17] A. Klapuri, “Sound onset detection by applying psychoacoustic knowledge,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’99)*, 1999.