Audio-Video Analysis of Musical Expressive Intentions

Ingrid Visentini[1], Antonio Rodà[1], Sergio Canazza[2], and Lauro Snidaro[1]

[1] University of Udine, Dept. of Mathematics and Computer Science, via Margreth 3, 33100 Udine, Italy
[2] University of Padova, Dept. of Information Engineering, Via Gradenigo 6/B, 35131 Padova, Italy
{ingrid.visentini,antonio.roda,lauro.snidaro}@uniud.it, canazza@dei.unipd.it

Abstract. This paper presents a preliminary study on the relation between audio-video streams and high-level information related to expressive nuances. A violinist was asked to play three musical excerpts several times, each one inspired by one of nine different expressive intentions. Perceptual tests was carried out using both audio-only and audio-visual recordings of the performances. The results demonstrate that the visual component aids the subjects to better recognize the different expressive intentions of the musical performances, showing that the fusion of audio-visual information can significantly improve the degree of recognition given by single means.

1 Introduction

This paper presents a preliminary study on the relation between audio-video streams and high-level information related to expressive nuances. At the moment, our concern is focused on audio-visual recordings of musical performances, as they have interesting applications both in multimedia information retrieval and in performing art contexts.

The sharing of increasingly large digital audio-visual libraries of musical performances over the network demands sophisticated tools to enable users to easily find the requested content. The textual approach used by today's search engines has limitations in its application to audio-visual files, because it allows only searching by metadata (i.e., title, author, genre, and so on), not by content. So if metadata, which are usually added manually, are incorrect or do not match with the content, the search can fail. Moreover, the user may not know exactly what document he/she is looking for, but might want to browse the audio-visual library to search for a musical performance that meets certain criteria: for example, a relaxing content or "something hard". In recent years, much progress has been made toward developing tools for content-based retrieval in audio-only documents (see [6] and [10] for a review). One of the most used approaches is to define a set of features that describe certain characteristics of sound and can be used to automatically classify the songs according to a determined list of categories. Almost completely unexplored is the joint use of audio and video analysis to improve the classification task.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 219–228, 2011.

Much contemporary music can take advantage of multi-modality to enhance performance as a globally engaging experience: music can be considered a conveyor of expressive content related to performance gestures. Several audio-visual operas (e.g. "Medea" by Adriano Guarnieri) explicitly insist on wanting to achieve an expressive matching between instrumental gestures and physical movement, so that both gestures would reinforce each other producing a more powerful and complete message. Although our work is at a basic research level, we believe that a deeper understanding of the relations among musical and video stimuli may improve the design of multisensorial interfaces, towards an effective mediation technology for music creation/production and content access/fruition.

In particular, this paper aims to study the gestures in relation of musical performances inspired by different *expressive intentions* [3]. This term refers to the expressive nuances that a musician wants to convey by means of its performance and includes emotions, affects as well as other sensorial aspects of a gesture. The relation between music and emotions has been largely investigated by the scientific community (see [7] and [8] for review). Mion & De Poli [9] asked three musicians to play several times a few short melodies, following different expressive intentions described by a set of affective and sensorial adjectives. A set of features, considered to be particularly representative of the expressive nuances of the performances, were extracted on the base of a frame size of 4 seconds. Results showed that, using the selected features, a linear classifier can recognize the expressive intentions of the songs, with an accuracy better than chance. Not many, however, are the studies that analyse the movements related with the expressive intentions in music. Camurri et al. [2] defined a multi-layer model to represent common characteristics of different sensorial domains, such as sounds and physical gestures. Dahl and Friberg [5] sudied the role of the different body parts in conveying emotional intentions during music performance, finding that head movement plays an important role in the communication of emotional content. Castellano et al. [4] was asked a pianist to play the same excerpt with different emotionally expressive intentions. The body movements captured by a camera positioned above the performer were analyzed via an automated system capable of detecting the temporal profiles of two motion cues: the quantity of motion of the upper body and the velocity of head movements. Results showed that both were sensitive to emotional expression, especially the velocity of head movements.

This paper has two objectives: i) to verify if an audio-video stream allows to recognize the performer's expressive intentions better than the audio-only stream; if so, ii) to find a set of descriptors of the video stream, to be related with the expressive intentions. We have recorded some musical excerpts and calculated statistics on both human perception (Section 2.3) and video sequences (Section 2.4). Since our target applications are the libraries of audio-visual and live art performances where we general can not have a control of the shooting condition, we chose to record in a poorly controlled environment, in terms of lighting and viewpoint, and without using markers. This choice has obviously affected the definition of the features to be extracted. As far as it regards video, we extract SURF and Lukas-Kanade features to determine movements and speed of the violinist playing the 27 musical excerpts.

2 Perceptual Experiments

We carried out two perceptual experiments to verify how the recognition of expressive intentions change in the following conditions: 1) audio-only musical stimuli are presented; 2) the audio stimuli are associated to a video of the musician playing the musical excerpts. The underlying assumption is that the video component makes it easy to better discriminate the expressive content of the musical excerpts.

2.1 Material

A violinist was asked to play three musical excerpts several times, each one inspired by one of the expressive intentions described by the following adjectives: happy, sad, angry, calm, hard, soft, heavy, light, and normal. The adjectives were chosen among the most widely used in studies of music performance: four refer to the emotional domain and four to the sensorial one. The normal performance, i.e. a performance that lacks a specific expressive intention, was introduced as a term of comparison to better assess the changes induced by the other expressive intentions. The three musical excerpts were chosen to represent different musical genres: a piece belonging to the Western classical repertoire (the incipit of the Violin Sonata Op. 1 No. 12 by G. F. Haendel), a popular melody (*Twinkle Twinkle Little Star*), and a jazz standard (*I Got Rhythm* by G. Gershwin). The performances were captured by one microphone and the audio signal were recorded in monophonic digital form at 24 bits and 48000 Hz.

Moreover, the musical performances was recorded also by two digital cameras observing the performer form different view angles. The two videos, synchronized with the audio track, have been processed to extract features that could capture the movement of the performer. The idea is then to extract the same type of features from each musical performance and for each expressive variation of it. The final goal is to compare the quantities computed for each variation and see if significant differences are observable that could lead to a discrimination of the expressive intentions of the performer.

The videos were acquired in non-constrained conditions of background and illumination. We extracted and analyzed Lukas-Kanade and SURF features from both videos to capture the movements of the performer.

2.2 Method

The procedure follows the one already used by Bigand in [1]. The experiment was conducted using an especially developed software interface. Participants were presented with a visual pattern of 27 loudspeakers, representing the 27 excerpts in a random order. They were required first to listen to all of these excerpts and to focus their attention on the expressive intention of the excerpts. They were then asked to look for excerpts with similar expressive intention and to drag the corresponding icons in order to group these excerpts. They were allowed to listen to the excerpts as many times as they wished, and to regroup as many excerpts as they wished. Both the experiments were performed by a total of 40 participants. Of these, 20 did not have any musical experience and are referred to as non-musicians and 20 have been music students for at least five years are referred to as musicians.

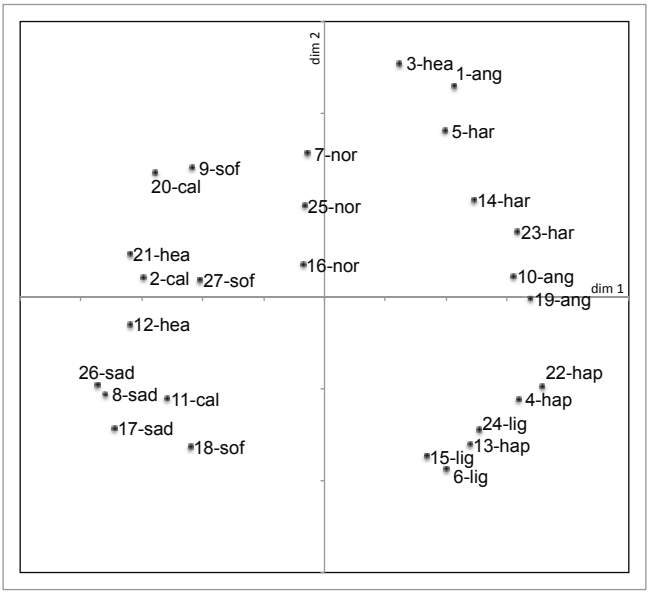

Fig. 1. Multi-Dimensional Scaling of the subjects' answers in the audio-only test

2.3 Results

The excerpts were numerated as follow: from 1 to 9 the performances of the *Violin Sonata* in the order angry, calm, happy, hard, heavy, light, normal, sad, and soft; from 10 to 18 *Twinkle Twinkle Little Star* with the same order; form 19 to 27 *I Got Rhythm*. Participants have formed an arbitrary number N of groups, named G_k. Each group contains the stimuli that the a subject thinks are characterized by the same or a similar expressive intention. The dissimilarity matrix A is defined by counting how many times two excerpts i and j are not included in the same group.

Table 1. Average distance measured between expressive intentions in the audio-only test

	ang	cal	hea	hap	har	lig	nor	sad	sof
ang	28.3	38.7	35.2	29.0	**27.8**	33.2	35.4	38.9	38.4
cal	38.7	**26.7**	32.1	38.7	37.9	37.7	33.7	29.6	27.4
hea	35.2	32.1	33.0	37.9	35.1	37.0	33.7	**30.1**	33.8
hap	29.0	38.7	37.9	**23.0**	32.1	28.2	35.9	39.0	37.9
har	27.8	37.9	35.1	32.1	**27.3**	32.9	33.8	39.1	36.7
lig	33.2	37.7	37.0	28.2	32.9	**26.3**	33.7	38.8	36.8
nor	35.4	33.7	33.7	35.9	33.8	33.7	**27.7**	36.8	31.6
sad	38.9	29.6	30.1	39.0	39.1	38.8	36.8	**21.7**	30.3
sof	38.4	**27.4**	33.8	37.9	36.7	36.8	31.6	30.3	30.3

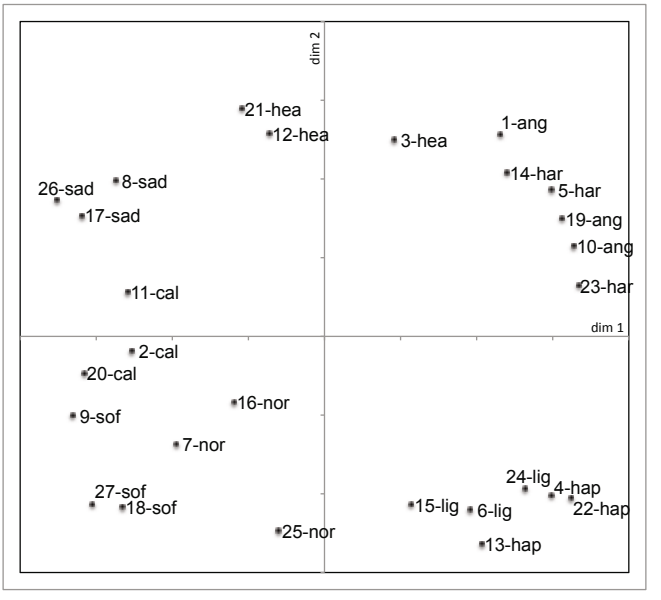

Fig. 2. Multi-Dimensional Scaling of the subjects' answers in the audio-video test

Table 2. Chi-square test on the audio-only test

	ang	cal	hea	hap	har	lig	nor	sad	sof
p-value	0.057	0.023	0.370	5.2e-04	0.038	0.011	0.054	4.4e-05	0.268

I.e., $\forall i, j = 1, ..., 27$ and $\forall k = 1, ..., N$

$$A[i,j] = \begin{cases} A[i,j] + 1 \text{ if} i \in G_k \wedge j \notin G_k \\ A[i,j] \qquad \text{otherwise} \end{cases} \tag{1}$$

Experiment 1: Audio-only Stimuli. The dissimilarity matrix from the experiment 1 was analysed by means of a Multi-Dimentional Scaling (MDS) method. The location of the 27 excerpts along the two principal dimensions is represented in Figure 1. The excerpts that are close in this space are those evaluated to be more similar (in terms of expressive characteristics) by the subjects. It can be noted the three normal performances located in the middle of the two-dimensional space, a cluster composed by the hard and angry performances in the upper right quadrant, a cluster with the happy and light performances in the lower right quadrant, and the three sad performances in the lower left quadrant. On the contrary, heavy and soft performances are not clustered, meaning the subjects did not recognize correctly these expressive intention.

Table 1 shows the average values calculated by grouping the entries of the dissimilarity matrix with the same expressive intention. The performances calm have, among

Table 3. Average distance measured between expressive intentions in the audio-video test

	ang	cal	hea	hap	har	lig	nor	sad	sof
ang	**19.7**	38.7	31.9	33.8	25.3	34.1	37.4	38.8	38.8
cal	38.7	**21.7**	33.0	39.0	38.7	34.7	31.2	31.7	31.3
hea	31.9	33.0	**24.3**	37.7	32.7	38.1	34.7	33.4	35.0
hap	33.8	39.0	37.7	**20.3**	34.1	25.0	35.4	37.6	38.1
har	25.3	38.7	32.7	34.1	**23.7**	33.6	36.3	38.8	38.3
lig	34.1	34.7	38.1	25.0	33.6	**22.3**	33.6	36.9	36.6
nor	37.4	31.2	34.7	35.4	36.3	33.6	**29.3**	34.2	30.8
sad	38.8	31.7	33.4	37.6	38.8	36.9	34.2	**18.7**	33.0
sof	38.8	31.3	35.0	38.1	38.3	36.6	30.8	33.0	**19.7**

Table 4. Chi-square test on the audio-video test

	ang	cal	hea	hap	har	lig	nor	sad	sof
p-value	6.7e-06	9.0e-05	1.9e-03	9.2e-06	8.3e-04	7.7e-04	0.122	3.1e-07	2.7e-06

them, a average dissimilarity of 26.7, which is smaller than the average dissimilarities between calm and the other expressive intentions. The same is true for the expressive intentions happy, hard, light, and sad. In all these cases, a Chi-square test (see Table 2) showed that these values are statistically significant ($p < 0.05$). The subjects' responses regarding the expressive intentions angry, heavy, normal, and soft, instead, are not significant.

Experiment 2: Audio-Video Stimuli. While the results of the experiment 1 show that some expressive intentions are confused, Figure 2 and Tables 3 and 4 show that all the expressive intentions are properly discriminated in the experiment with audio-visual stimuli.

2.4 Video Analysis

An example of feature extraction in video sequences is presented in 3, where the performer movements are followed by the Lukas-Kanade features in the first and second view (Figure 3 (a) and (b) respectively), and the features motion accumulator after 100 frames (in Figure 3 (c) and (d)).

We have dumped the values of the position of each feature for each frame and computed the following statistics:

$$MeanSpread = \frac{1}{n}\sum_{i=1}^{n} \max\left(d_i(x,y)\right) - \min\left(d_i(x,y)\right) \tag{2}$$

$$MeanDerivative = \frac{1}{nT}\sum_{t=1}^{T}\sum {}^{n}_{i=1} |d'_{i,t}(x,y)| \tag{3}$$

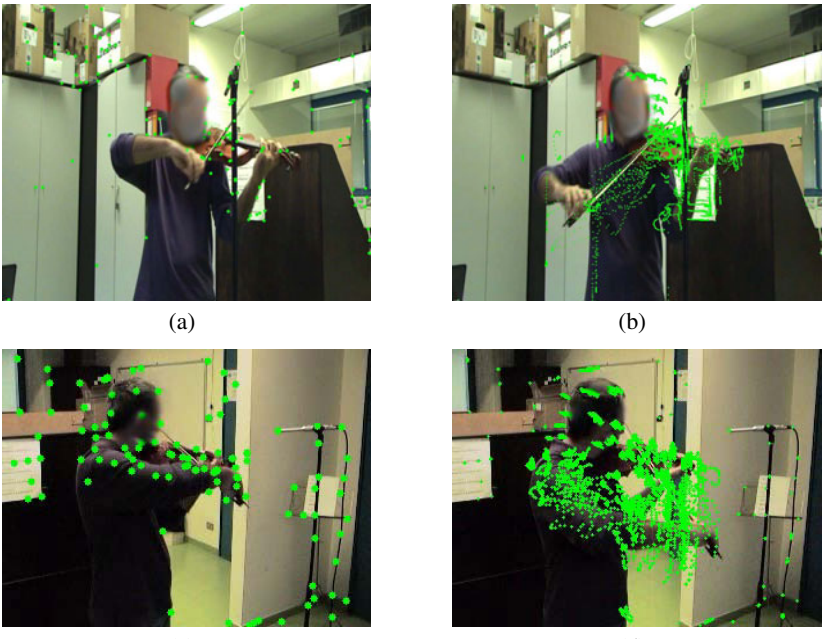

Fig. 3. Examples of frames taken from video sequence, picturing the performer dotted with the Lukas-Kanade features in the first view (a), the features motion accumulator at frame 100 (b), and their behaviour in the second view (c) with the accumulation of their motion after 100 frames in (d)

The first one (2) indicates the average distance covered by the features, calculated as the difference from the maximum and minimum position magnitude. The average is computed on all the features, while the magnitude of the i-th feature is given by

$$d_i(x, y) = \sqrt{x^2 + y^2} \tag{4}$$

where (x, y) is the position of the feature in the image coordinates system. The mean derivative in (3) indicates the average speed of the features. Notice that many features can be stationary since they may be positioned in background regions. For this reason we have not included in the statistics above the features whose "movement" is below a pre-set threshold of 10. In the equations 2 and 3 index i runs from 1 to n over the features in a single frame, while t, where $1 \leq t \leq T$, runs over the total number of frames of the video.

We repeated the test for each musical excerpt, and we obtained the results illustrated in Figure 4 and 5 for first and second camera respectively. In both figures, on the x axis there are the expressive intentions, while on the y axis the mean spread (top graph) and mean derivative (bottom graph) of the features extracted from the video sequence. The standard deviation of the two measures is depicted with an error bar.

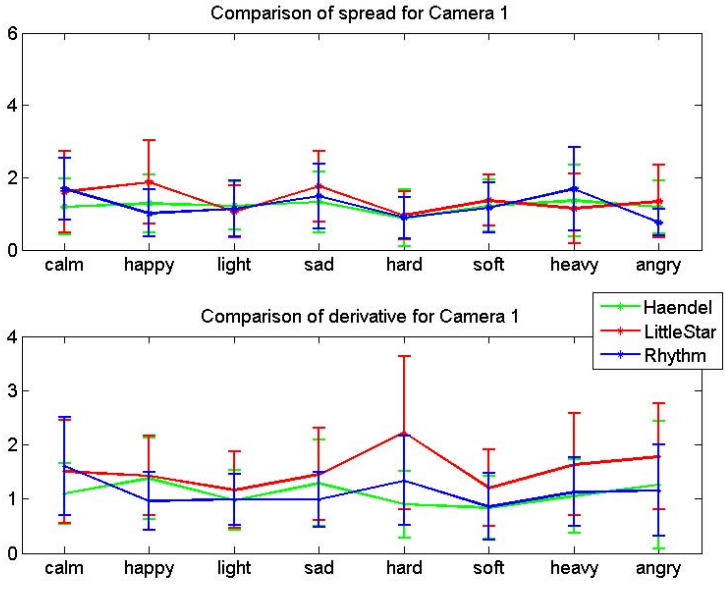

Fig. 4. Mean spread (top) and mean derivative (bottom) of the features extracted from the first camera

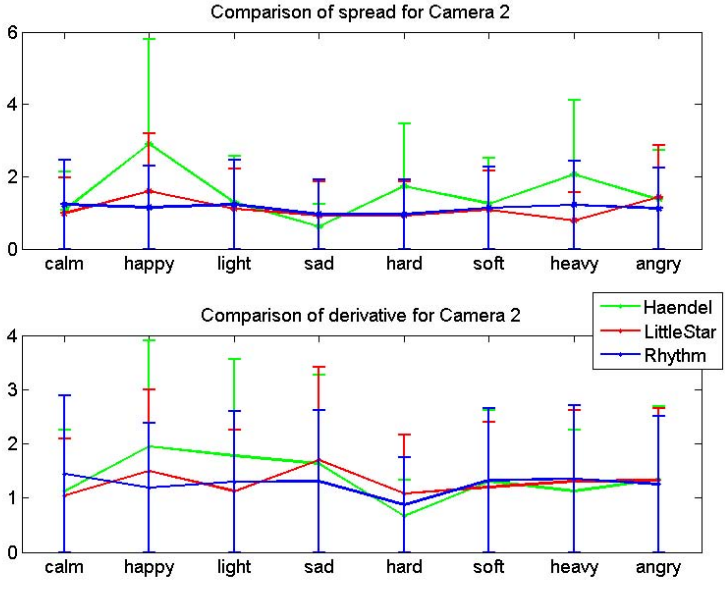

Fig. 5. Mean spread (top) and mean derivative (bottom) of the features extracted from the second camera

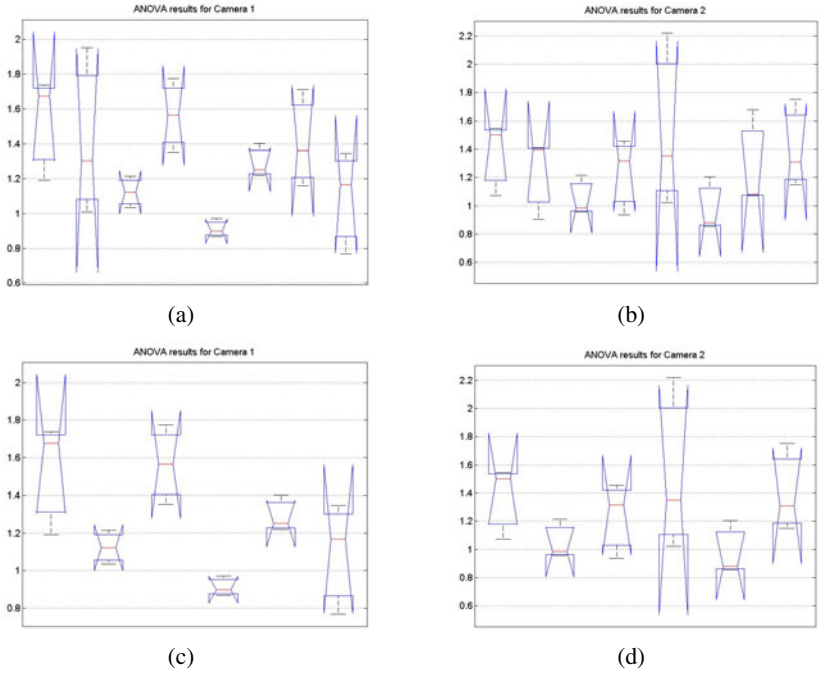

Fig. 6. Results of the ANOVA test for first (a) and second (b) camera on the full expression intentions set. In the second row, the outcomes with a reduced set are presented for the two cameras respectively

We used a one-way ANOVA to test if the expressive intentions are different and separable, considering as null hypothesis the equality of their means. Considering the full set of intentions, we obtained $F(7, 16) = 2.32$ and $p = 0.0773$ for the first camera and $F(7, 16) = 0.9$ and $p = 0.5330$ for the second. Comparing these values with the F-ratio table references, the hypothesis has high probability of being accepted, that is the intentions are not separable. Discarding two expressive intentions with the larger standard deviation, that are *happy* and *heavy* for both cameras, the new values were $F(5, 12) = 4.99$ and $p = 0.0106$ for the frontal view and $F(5, 12) = 1.22$ and $p = 0.3599$ for the other. It is clear that in the case of frontal camera, the removal of the classes with large standard deviation allows to separate the means of remaining expression intentions with a high probability, but in the case of a side camera the situation does not gain much benefit from discarding some observations. Another consideration is that *heavy* and *happy* turn out to be ambiguous expression after automatic video analysis. In Figure 6 are presented illustrations after ANOVA test. In the left column the results for the first camera with all the intentions and with the above mentioned subset are presented. In the second column, the results for the second camera are shown, and the classes are not separable even after the exclusion of the two most uncertain classes.

3 Conclusions

In this paper we presented an analysis of human perception applied to musical expressive intention recognition. The results demonstrate that the visual component aids the subjects to better recognize the different expressive intentions of the musical performances. As humans were tested with audio only, and then with the fusion of audio and video, we selected video features to obtain a fair comparison. We extracted SURF and Lukas-Kanade features to determine movements and speed of the violinist playing 27 musical excerpts. To summarize, the fusion of audio-visual information can significantly improve the degree of expression intention recognition given by single means. Future research direction will be oriented to fuse audio and video features automatically extracted from sequences and to compare them with the results of human recognition.

References

1. Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., Dacquet, A.: Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition and Emotion 19(8), 1113–1139 (2005)
2. Camurri, A., De Poli, G., Leman, M., Volpe, G.: Communicating expressiveness and affect in multimodal interactive systems. IEEE Multimedia 12(1), 43–53 (2005)
3. Canazza, S., De Poli, G., Rodà, A.: Analysis of expressive intentions in piano performance. Journal of ITC Sangeet Research Academy 16, 23–62 (2002)
4. Castellano, G., Mortillaro, M., Camurri, A., Volpe, G., Scherer, K.: Automated analysis of body movement in emotionally expressive piano performances. Music Perception 26(2), 103–119 (2008)
5. Dahl, S., Friberg, A.: Visual perception of expressiveness in musician's body movements. Music Perception 24, 433–454 (2007)
6. Downie, J.S.: Music information retrieval. Annual Review of Information Science and Technology 37, 295–340 (2003)
7. Juslin, P.N., Sloboda, J.A.: Music and emotion. Theory and research. Oxford University Press, Oxford (2001)
8. Kirke, A., Miranda, E.R.: A survey of computer systems for expressive music performance. ACM Computing Surveys 42(1) (2009)
9. Mion, L., De Poli, G.: Score-independent audio features for description of music expression. IEEE Trans. Speech, Audio, and Language Process. 16(2), 458–466 (2008)
10. Orio, N.: Music retrieval: A tutorial and review. Foundations and Trends in Information Retrieval 1(1), 1–90 (2006)