Towards a Gesture-Sound Cross-Modal Analysis

Baptiste Caramiaux, Frédéric Bevilacqua, Norbert Schnell

Towards a Gesture-Sound Cross-Modal Analysis

Baptiste Caramiaux, Frédéric Bevilacqua, and Norbert Schnell

Real Time Musical Interactions Team
IRCAM, CNRS - STMS,
1 Place Igor Stravinsky, 75004 PARIS, France
{baptiste.caramiaux,frederic.bevilacqua,norbert.schnell}@ircam.fr

Abstract. This article reports on the exploration of a method based on canonical correlation analysis (CCA) for the analysis of the relationship between gesture and sound in the context of music performance and listening. This method is a first step in the design of an analysis tool for gesture-sound relationships. In this exploration we used motion capture data recorded from subjects performing free hand movements while listening to short sound examples. We assume that even though the relationship between gesture and sound might be more complex, at least part of it can be revealed and quantified by linear multivariate regression applied to the motion capture data and audio descriptors extracted from the sound examples. After outlining the theoretical background, the article shows how the method allows for pertinent reasoning about the relationship between gesture and sound by analysing the data sets recorded from multiple and individual subjects.

Key words: Gesture analysis, Gesture-Sound Relationship, Sound Perception, Canonical Correlation Analysis

1 Introduction

Recently, there has been an increasing interest in the multimodal analysis of the expression of emotion as well as expressivity in music. Several works reveal that motor expression components like body gestures are always accompanying other modalities [23]. For instance, human face-to-face communication often combines speech with non-verbal modalities like gestures. In this context, multimodal analysis reveals co-expressive elements that play an important role for the communication of emotions. In a similar way, we'd like to explore the relationship between gestures and sound in the context of music performance and listening.

We are particularly interested in the relationship between sound and the movements of an individual or a group in a listening situation as well as the movements of a music performer that are related primarily to the production of sound, in addition to the musical intention and the expression of emotion ([16]).

In our current project, we develop a set of methods for the analysis of the relationship between different aspects of gestures and sound. We would like to be able to apply these methods to a variety of contexts, covering the performance

of traditional and electronic (virtual) instruments as well as different music listening scenarios. The goal of this work reaches the creation of tools for the study of gesture in musical expression and perception. In a greater context, these tools contribute to the development of novel paradigms within the intersection between music performance and music listening technologies.

In this paper, we present a new approach to the quantitative analysis of the relationship between gesture and sound. The article is organized as follows. We first present a review of related works. Then we introduce in section 3 the multivariate analysis method called canonical correlation analysis. In section 4 we present the experimental context including our data capture methods and we show results on feature selection and correlation analysis of collected data. We discuss these results in 5. Finally, we conclude and give the implications on further works in section 6.

2 Related Work

The concept of embodied cognition has been adopted by a wide community of researchers. In this context, the relationship between gesture and sound has come into interest to interdisciplinary research on human communication and expression.

Some recent researches in neurosciences ([13], [25]) and others in perception ([26], [2], [18]) have shown that action plays a predominant role in perception insisting on the inherently multimodal nature of perception. In [12], [14], [1] the authors show that gesture and speech are to some extent complementary co-expressive elements in human communication.

Research in the domain of music and dance has studied the embodiment of emotion and expressivity in movement and gesture. Leman ([16]) has widely explored various aspects of music embodiment based on the correlation between physical measurements and corporeal articulations in respect to musical intention. Camurri et al. in [4] show that emotion can be recognized in a dancing movement following dynamic features such as *quantity of motion* extracted from motion capture data. Dahl et al. in [5] show to what extent emotional intentions can be conveyed through musicians' body movements. Moreover, Nusseck and Wanderley in [19] show that music experience is multimodal and is less depend on the players' particular body movements than the player's overall motion characteristics.

Several recent works have studied gestures performed while listening to music revealing how an individual perceives and imagines sound and sound production as well as music and music performance. In [6], [10] and [7] the authors explore the relationship between gesture and musical sound using qualitative analysis of the gestural imitation of musical instrument performance (*air-instruments*) as well as free dance and drawing movements associated with sounds (*sound-tracing*). For instance, [6] shows that air-instrument performance can reflect how people perceive and imagine music highly depending on their musical skills.

On the other hand, only a few works have taken a quantitative approach and are mostly focussing on the synchronisation between gestures and music. In [15], Large proposes a pattern-forming dynamical system modelling the perception of beat and meter that allows for studying the synchronisation and rhythmic correspondence of movement and music. Experiments in which subjects were asked to tap along with the musical tempo have revealed other pertinent characteristics of the temporal relationship between movement and music ([22], [17], [24]) such as negative asynchrony, variability, and rate limits. In [17], the authors give a quantitative analysis of the ensemble musicians' synchronization with the conductor's gestures. The authors have used cross-correlation analysis on motion capture data and beat patterns extracted from the audio signal to study the correspondence between the conductor's gestures and the musical performance of the ensemble. Lastly, Styns ([24]) has studied how music influences the way humans walk analysing the correspondence between kinematic features of walking movements and beat patterns including the comparison of movement speed and walking tempo in addition to the analysis of rhythmic synchronicity. He shows that walking to music can be modelled as a resonance phenomenon (with resonance frequency at 2Hz).

In our work we attempt to introduce a method for the quantitative multimodal analysis of movement and sound that allows for the selection and analysis of continuous perceptively pertinent features and the exploration of their relationship. It focuses on free body movements performed while listening to recorded sounds. The mathematical approach is a general multivariate analysis method that has not been used yet in gesture-sound analysis, but that has given promising results in the analysis of multimedia data and information retrieval ([11]).

3 Canonical Correlation Analysis: an Overview

Proposed by Hotelling in [9], Canonical Correlation Analysis (CCA) can be seen as the problem of measuring the linear relationship between two sets of variables. Indeed, it finds basis vectors for two sets of variables such that the correlations between the projections of the variables onto these basis vectors are mutually maximised. Thus, respective projected variables are a new representation of the variables in directions where variance and co-variance are the most explained.

Let us introduce some notations: bold type will be used for matrices ($\mathbf{X}$, $\mathbf{Y}$, etc...) and vectors ($\mathbf{u}$, $\mathbf{v}$, etc...). The matrix transpose of $\mathbf{X}$ will be written as $\mathbf{X}^T$. Finally, an observation of a random variable $\mathbf{v}$ will be written as v_i at time i.

Consider two matrices $\mathbf{X}$ and $\mathbf{Y}$ where the rows (resp. columns) are the observations (resp. variables). $\mathbf{X}$, $\mathbf{Y}$ must have the same number of observations, denoted m, but can have different numbers of variables, denoted n_x resp. n_y. Then, CCA has to find two projection matrices, $\mathbf{A}$ and $\mathbf{B}$, such as

$$\max_{\mathbf{A},\mathbf{B}} \left[\mathrm{corr}\left(\mathbf{XA}, \mathbf{YB}\right)\right] \tag{1}$$

Here *corr* denotes the correlation operator between two matrices. Usually, the correlation matrix of a matrix $\mathbf{M}$ of dimension $m \times n$ is the correlation matrix of n random variables (the matrix columns $\mathbf{m}_1, ..., \mathbf{m}_n$) and is defined as a $n \times n$ matrix whose (i, j) entry is $\text{corr}(\mathbf{m}_i, \mathbf{m}_j)$. The correlation between two matrices is the correlation between the respective indexed columns. Therefore $\mathbf{XA}$ and $\mathbf{YB}$ must have the same number of variables. $\mathbf{A}$ and $\mathbf{B}$ are $n_x \times \min(n_x, n_y)$ and $n_y \times \min(n_x, n_y)$ matrices. Let h be one arbitrary variable index in $\mathbf{XA}$ (as in $\mathbf{YB}$), equation (1) can be written as finding $\mathbf{a}_h$ and $\mathbf{b}_h$, $\forall h = 1... \min(n_x, n_y)$, that maximize:

$$\text{corr}(\mathbf{Xa}_h, \mathbf{Yb}_h) \tag{2}$$

We remind the reader that the correlation coefficient between two random variables is computed as the quotient between the covariance of these two random variables and the square root of the product of their variance. Let denote $\mathbf{C}(\mathbf{X}, \mathbf{Y})$ the covariance matrix. It is a positive semi-definite matrix and can be written as

$$\mathbf{C}(\mathbf{X}, \mathbf{Y}) = \hat{\mathbb{E}}\left[\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}^T \begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\right] = \begin{bmatrix}\mathbf{C}_{xx} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{C}_{yy}\end{bmatrix}$$

Thus we can formulate the problem from equation (2) using the previous notations: find $\mathbf{A}, \mathbf{B}$ such that the following quotient is maximized

$$\text{corr}(\mathbf{Xa}_h, \mathbf{Yb}_h) = \frac{\text{cov}(\mathbf{Xa}_h, \mathbf{Yb}_h)}{\sqrt{\text{var}(\mathbf{Xa}_h)\,\text{var}(\mathbf{Yb}_h)}} = \frac{\mathbf{a}_h^T \mathbf{C}_{xy} \mathbf{b}_h}{\sqrt{\mathbf{a}_h^T \mathbf{C}_{xx} \mathbf{a}_h \mathbf{b}_h^T \mathbf{C}_{yy} \mathbf{b}_h}} \tag{3}$$

One can show that equation (3) leads to a generalized eigenproblem of the form (see [8]):

$$\mathbf{M}_1 \mathbf{v} = \lambda \mathbf{M}_2 \mathbf{v}$$

Efficient methods can be implemented to find interesting projection matrices. The key terms for an understanding of CCA are: *canonical weights* (coefficients in $\mathbf{A}$ and $\mathbf{B}$); *canonical variates* (projected variables, $\mathbf{XA}$ and $\mathbf{YB}$); *canonical function* (relationship between two canonical variates whose strength is given by the canonical correlation).

Interpreting canonical correlation analysis involves examining the canonical functions to determine the relative importance of each of the original variables in the canonical relationships. Precise statistics have not yet been developed to interpret canonical analysis, but several methods exist and we have to rely on these measures. The widely used interpretation methods are: canonical weights, canonical loadings and canonical cross-loadings. In this paper we use the second one because of its efficiency and simplicity. Canonical Loadings measure the simple correlation between variables in each set and its corresponding canonical variates, i.e. the variance that variables share with their canonical variates. Canonical Loadings are computed as:

$$\textit{Gesture loadings} : \mathbf{L}_G = \text{corr}(\mathbf{X}, \mathbf{U})$$

$$\textit{Sound loadings} : \mathbf{L}_S = \text{corr}(\mathbf{Y}, \mathbf{V})$$

4 Cross-Modal Analysis

We applied the method based on CCA to some examples of data collected in an experiment with subjects performing free body movements while listening to sound recordings imagining themselves producing the sound. Given the setup of the experiment, gesture and sound can be assumed as highly correlated without knowing their exact relationship that may be related to the subjects' sound perception, their intention of musical control, and their musical and motor skills. In this sense, the collected data sets have been a perfect context to explore the developed method and its capability to support reasoning about the relationship between gesture and sound.

4.1 Collected Data

The data has been collected in May 2008 in the University of Music in Graz. For the experiment 20 subjects were invited to perform gestures while listening to a sequence of 18 different recorded sound extracts of a duration between 2.05 and 37.53 seconds with a mean of 9.45 seconds. Most of the sound extracts were of short duration. Since the experience was explorative, the sound corpus included a wide variety of sounds: environmental and musical of different styles (classical, rock, contemporary).

For each sound, a subject had to imagine a gesture that he or she performed three times after an arbitrary number of rehearsals. The gestures were performed with a small hand-held device that included markers for a camera-based motion capture system recording its position in Cartesian coordinates. A foot-pedal allowed to synchronise the beginning of the movement with the beginning of the playback of the sound extract in the rehearsal as well as for the recording of the final three performances.

4.2 Gesture Data

As input of the analysis method, a gesture is a multi-dimensional signal stream corresponding to a set of observations. The most basic kinematic features are the position coordinates x, y, z, velocity coordinates v_x, v_y, v_z and acceleration coordinates a_x, a_y, a_z derived from the motion capture data. These features give a basic and efficient representation of postures and body movements describing their geometry and dynamics. For instance, Rasamimanana in [21] shows that three types of bow strokes considered in the paper are efficiently characterized by the features $(a_{\min}, a_{\max})$. In order to abstract from absolute position and movement direction, we calculate vector norms for position, velocity, and acceleration. To also consider movement trajectories, we additionally represent the gestures in an adapted basis using Frenet-Serret formulas giving *curvature* and *torsion* in the coordinate system $(\mathbf{t}, \mathbf{n}, \mathbf{b})$. In the same coordinate system, we add *normal* and *tangential accelerations* denoted by acc_N and acc_T (that replace previous acceleration).

Finally, at the input of the method a gesture is represented by a finite sequence of observations of the following variables:

$$\{position, velocity, acc_N, acc_T, curvature, radius, torsion\}$$

The CCA here permits to select the most pertinent features used in further calculations eliminating non-significant parameters.

4.3 Sound Features

The perception of sound has been studied intensively since one century and it is now largely accepted that sounds can be described in terms of their pitch, loudness, subjective duration and "timbre". For our exploration, we extract a set of audio descriptors from the audio files used in the experiment that have been shown to be perceptively relevant (see [20]). While we easily can rely on loudness and pitch the perceptual relevance of audio descriptors for timbre and its temporal evolution is less assured. Nevertheless, we have chosen to use *sharpness* corresponding to the perceptual equivalent to the spectral centroïd. Pitch has been discarded since in musical performance it generally requires high precision control associated to expert instrumental gestures (defined as *selection gestures* in [3]).

At the input of the method a sound is represented by a finite sequence of observations of the following variables:

$$\{loudness, sharpness\}$$

Their perceptual characteristic allows the easy interpretation of gesture-sound relationship analysis.

4.4 Results

For free body movements performed while listening to recorded sound extracts, we are interested in investigating how gesture can explain sound through sound features and how sound can highlight important gesture characteristics. Among the whole set of sounds we chose two: the sound of an ocean wave and a solo flute playing a single note with strong timbre modulation (extract from *Sequenza I* for flute (1958), by Luciano Berio). These two sounds appeared to be the most pertinent extracts given the selection of audio descriptors discussed in 4.3. The set of two perceptual audio descriptors computed on each sound can be seen in figure 1.

The wave sound is characterized by a spectral distribution similar to a white noise passing through a specific filter. It leads to a sharpness feature highly correlated with the loudness (correlation coefficient of 0.814). Since the flute example characteristic resides in a continuous transformation of its spectrum without significatively changing the fundamental frequency, its computed loudness and sharpness are less correlated (its correlation coefficient is -0.61).

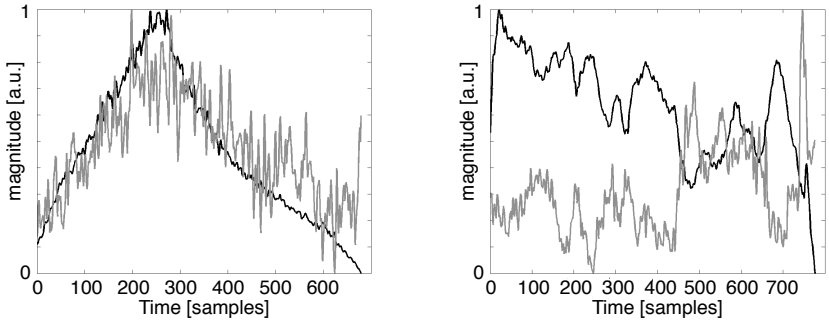

Fig. 1. Loudness and Sharpness. On the left, feature values are plotted for the wave sound. The line corresponds to loudness, and the gray line sharpness. The same features for the flute timbre example are plotted on the right.

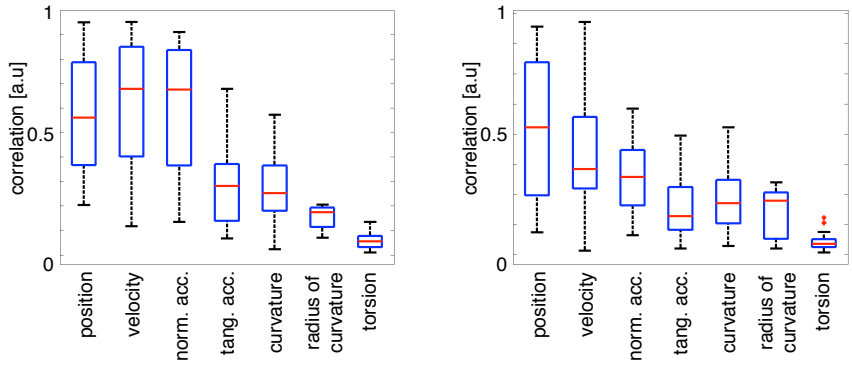

Fig. 2. Relevant gesture parameters. Each parameter is analysed together with the audio descriptors using CCA for 42 gestures. Results for the wave sound are plotted on the left side while flute results can be seen on the right.

First, gesture parameters considered as pertinent in the context cannot be chosen arbitrarily. Our analysis method can be applied to select a subset of pertinent gesture parameters using one gesture and many audio descriptors. In this way, the method operates as a multiple regression: the gesture parameter is predicted from audio descriptors. Each analysis returns one correlation coefficient corresponding to the canonical function strength between the current gesture parameter and the audio canonical component. 42 gestures are considered landing 42 canonical analysis iterations for each gesture parameter and each sound. Figure 2 shows two box plots corresponding to this process as applied to the wave and flute sounds. Three principal features are emphasized: position (index 1),

velocity (index 2), and normal acceleration (index 3). Since these features have the highest correlation means among those in the set of gesture parameters, they constitute a set of pertinent parameters related to the wave and flute sounds. Nevertheless, selection based on correlation means returns more significant results for the wave sound. For both cases, torsion has been discarded because the data derived from the motion capture recordings were very noisy.

Therefore, the selected subset of gesture parameters is $\{position, velocity, acc_N\}$. Canonical correlation analysis has been used as a selection tool; now we apply this method in our search for the intrinsic relationships between the two sets of data. In the first step, we discard outliers related to the first and the second canonical component. This leads to two subsets: 14 gestures among 42 for the wave example and 10 gestures for the flute example. Following the previous notations, CCA returns two projection matrices $\mathbf{A}, \mathbf{B}$ whose dimensions are 3×2 and 2×2 for each gesture, respectively. Loadings are computed at each step; figure 3 and 4 illustrate their statistics. The figures show the variance shared by each original variable with its canonical component for all gestures. Canonical gesture loadings are on the left side of the figures while audio descriptors respective canonical loadings are on the right. The first component is placed above the second one.

The wave case is illustrated by figure 3 and can be interpreted as follows. Gesture parameter velocity and normal acceleration are the most represented in the first canonical component: around 90% of their variance is explained. In the audio space, one original variable is clearly highlighted: the loudness (at the top of figure 3). In other words, the first canonical function correlates $\{velocity, acc_N\}$ to $\{loudness\}$.

Position contributes the most to the second canonical component in the gesture space while the sharpness descriptor is predominant in this case. So second canonical function correlates $\{position\}$ to $\{sharpness\}$ (at the bottom of figure 3).

One can remark that analysis reveals that loudness and sharpness descriptors can be separated when considering sound with gesture while they were highly correlated (figure 1).

A similar interpretation can be given for the flute timbre sound showed in figure 4. In this case, we have:

$$\begin{aligned} \text{first function} &: \quad \{position\} \rightarrow \{loudness\} \\ \text{second function} &: \{velocity, acc_N\} \rightarrow \{sharpness\} \end{aligned}$$

5 Discussion

To analyse the cross-modal relationship between gesture and sound, a multivariate analysis method is used in two ways: first for the selection of pertinent gesture features, then for the analysis of the correlation between the selected features with the audio descriptors. In the first step, the selection yields a subset

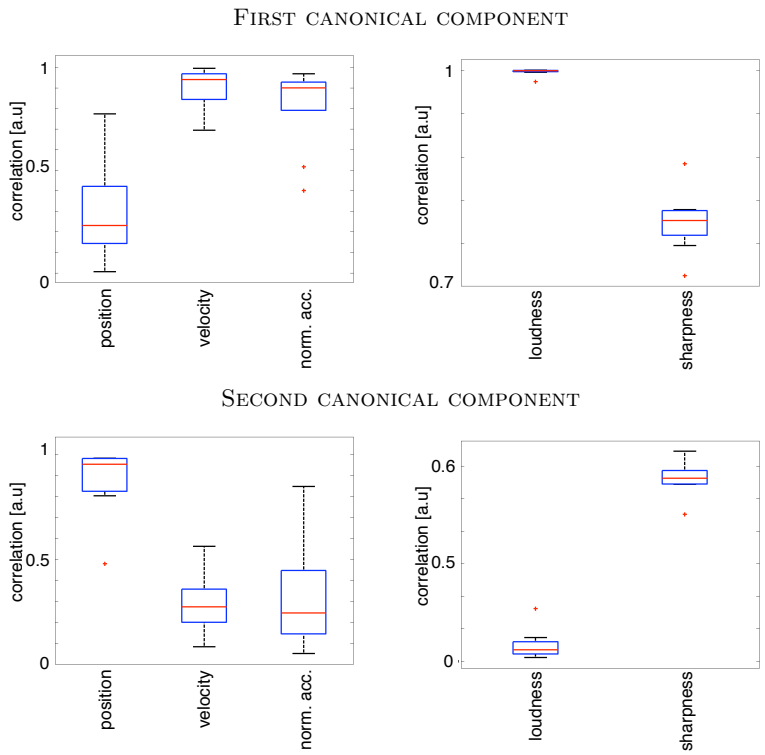

Fig. 3. Canonical loadings for the wave sound. Each row is a canonical component. Gesture parameter loadings are plotted on the left while audio descriptors can be seen on the right. Top: *velocity* and acc_N are correlated to *loudness*. Bottom: *position* is correlated to *sharpness*.

of movement features that best correlate with the audio descriptors. The low correlations obtained for some of the features have been discarded for further exploration. This seems to be coherent with kinematic studies of human gestures:

- Tangential acceleration is the acceleration component which is collinear to the velocity vector. If we consider the two-thirds power law by Viviani and Flash ($A = K.C^{2/3}$ where A is the angular velocity, C the curvature and K a constant), normal acceleration is related to curvature by $acc_N = K'.C^{1/3}$, where K' is a constant. In this case, tangential acceleration does not convey relevant information.
- The fact that curvature is no longer pertinent means there is no linear relation either between curvature and the audio descriptors or between the curvature and other gesture parameters. This result is in agreement with the previous kinematic law and can be also applied to the radius of curvature.

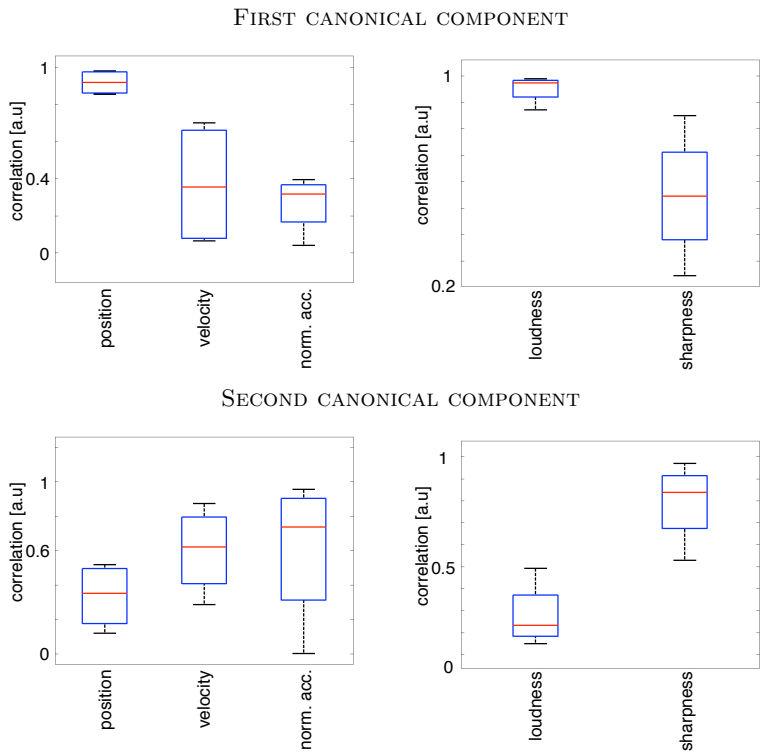

Fig. 4. Canonical loadings for the flute sound. Each row is a canonical component. Gesture parameter loadings are plotted on the left while audio descriptors can be seen on the right. Top: *position* is correlated to *loudness*. Bottom: *velocity* and acc_N are correlated to *sharpness*.

The next step of the analysis explores the correlation of selected movement features with the audio descriptors. The results of this analysis are correlations highlighting pertinent aspects of the gesture-sound relationship. Without surprise the subjects seem to favour gestures correlating with perceptual audio energy (*loudness*).

In the case of the wave sound, velocity or normal acceleration are highly correlated to loudness. Confronting this result with performance videos, one can see that the subjects are concerned about sonic dynamics and continuity. Increasing audio energy implies increasing velocity, i.e. increasing kinetic energy. Here the analysis reveals that the subjects tend to embody sound energy through the energy of their movement.

On the other hand, for the gestures performed on the flute sound we observe a high correlation between the norm of the position and the loudness. Instead of embodying the sound dynamic the subjects rather tend to transcribe its temporal evolution tracing the modulation of the sound feature over time. As the variation

of audio energy in the flute example is rather subtle compared to the wave sound, the subjects seem to adapt their strategy for the imagined sound control.

At last, we have started to inspect data of particular subjects that may reveal individual strategies and skills. For instance, considering the velocity feature, defined as $velocity^2 = v_x^2 + v_y^2 + v_z^2$, one can bring directional information to the analysis splitting $velocity^2$ into three specific variables: v_x^2, v_y^2, v_z^2. Canonical correlation analysis is no longer constrained to a uniform weight equal to 1 in the resulting linear combination but finds an optimal set of weights favouring directions. In other words, the analysis method takes into account the movement asymmetries. For the selection of movement parameters among a redundant set of extracted features, a trade-off has to be found between achieving a complete description of the movement and avoiding redundancies.

6 Conclusion and Future Works

Our goal was to study the relationship between gesture and sound. Gesture was considered as a set of kinematic parameters representing a free movement performed on a recorded sound. The sound was considered as a signal of feature observations. The method used in the paper arises from multivariate analysis research and offers a powerful tool to investigate the mutual shared variance between two sets of features. Objective results inferred from the application of CCA as a selection tool was presented. In addition, more subjective conclusions concerning mapping from the gesture parameter space to the audio descriptor space was highlighted. Thereby, we saw in this paper that gestural expression when relating to sounds can be retrieved considering gesture-sound as a pair instead of as individual entities.

However, the method suffers from some restrictive limitations. First of all, canonical functions correspond to linear relations so CCA cannot exhibit non-linear relations between variables. Besides, since we must restrict the variable sets to finite sets that encode only a part of the information contained in both gestures and sounds, the correlation (i.e. variance) as an objective function is not always relevant when real signals are analysed. The correlation involved in CCA could be replaced by the mutual information. By arising the statistical order of the multivariate relation, the main idea is to find canonical variates that are maximally dependent. It should lead to a more complete semantic interpretation of gesture-sound relationships in a musical context. To summarize, the method presented in this paper has given promising results and further works will consist in refining the method using information theory.

7 Acknowledgments

This work was supported by the ANR project 2PIM/MI3. Moreover, we would like to thank the COST IC0601 Action on Sonic Interaction Design (SID) for their support in the short-term scientific mission in Graz.

References

1. Bergmann, Kirsten and Kopp, Stefan. Co-expressivity of speech and gesture: Lessons for models of aligned speech and gesture production. *Symposium at the AISB Annual Convention: Language, Speech and Gesture for Expressive Characters*, pages 153–158, December 2007.
2. Berthoz, Alain. *Le Sens du mouvement.* Odile Jacob, Paris, France, 1997.
3. Cadoz, Claude and Wanderley, Marcelo M. Gesture-music. In *Trends in Gestural Control of Music*, pages 1–55. Ircam, Paris, France, 2000.
4. Camurri, Antonio, Lagerlöf, Ingrid, and Volpe, Gualtiero. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1-2):213–225, July 2003.
5. Dahl, Sofia and Friberg, Anders. Expressiveness of musician's body movements in performances on marimba. *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop, GW 2003*, LNAI 2915:479–486, 2003.
6. Godøy, R. I., Haga, E., and Jensenius, A. R. Playing "air instruments": Mimicry of sound-producing gestures by novices and experts. *Gesture in Human-Computer Interaction and Simulation, 6th International Gesture Workshop, GW 2005*, 3881/2006:256–267, 2005.
7. Haga, Egil. *Correspondences between music and body movement.* PhD thesis, University of Oslo, Department of Musicology, 2008.
8. Hair, Joseph F., Black, William C., Babin, Barry J., and Anderson, Rolph E. *Multivariate Data Analysis (7th Edition).* Prentice Hall, New Jersey, USA, February, 2009.
9. Hotelling, Harold. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, December 1936.
10. Jensenius, Alexander Refsum. *Action-Sound, Developing Methods and Tools to Study Music-Related Body Movement.* PhD thesis, University of Oslo, Department of Musicology, 2007.
11. Kidron, Einat, Schechner, Yov Y., and Elad, Michael. Pixels that sound. *IEEE Computer Vision & Pattern Recognition (CVPR 2005)*, 1:88–95, June 2005.
12. Kita, Sotaro and Asli, Özyürek. What does cross-linguistic variation in semantic coordination of speech and gesture reveal ?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32, 2003.
13. Kohler, Evelyne, Keysers, Christian, Umiltà, Alessandra, Fogassi, Leonardo, Gallese, Vittorio, and Rizzolatti, Giacomo. Hearing sounds, understanding actions : Actions representation in mirror neurons. *Science*, 297:846–848, 2002.
14. Kopp, Stefan and Wachsmuth, Ipke. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, March 2004.
15. Large, Edward W. On synchronizing movements to music. *Human Movement Science*, 19(4):527–566, 2000.
16. Leman, Marc. *Embodied Music Cognition and Mediation Technology.* Massachusetts Institute of Technology Press, Cambridge, USA, 2008.
17. Luck, Geoff and Toiviainen, Petri. Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis. *Music Perception*, 24(2):189–200, 2006.

18. Noë, Alva. *Action in Perception.* Massachusetts Institute of Technology Press, Cambridge, USA, 2005.
19. Nusseck, Manfred and Wanderley, Marcelo M. Music and motion - how music-related ancillary body movements contribute to the experience of music. *Music Perception*, 26:335–353, 2009.
20. Peeters, Geoffroy. A large set of audio features for sound description. *CUIDADO Project*, 2004.
21. Rasamimanana, Nicolas, Fléty, Emmanuel, and Bévilacqua, Frédéric. Gesture analysis of violin bow strokes. *Gesture in Human-Computer Interaction and Simulation, 6th International Gesture Workshop, GW 2005*, 3881:144–155, 2006.
22. Repp, Bruno Hermann. *Musical Synchronization*, pages 55–76. Music, motor control and the brain, Oxford University Press, e. altenmüller, m. wiesendanger, j. kesselring (eds.) edition, 2006.
23. Scherer, Klaus R. and Ellgring, Heiner. Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, 7(1):158–171, 2007.
24. Styns, Frederik, van Noorden, Leon, Moelants, Dirk, and Leman, Marc. Walking on music. *Human Movement Science*, 26(5):769–785, 2007.
25. Metzinger Thomas and Gallese Vittorio. The emergence of a shared action ontology: Building blocks for a theory. *Consciousness and Cognition*, 12(4):549–571, 2003.
26. Varela, Francisco, Thompson, Evan, and Rosch, Eleanor. *The Embodied Mind: Cognitive Science and Human Experience.* Massachusetts Institute of Technology Press, Cambridge, USA, 1991.