# Analyzing Sound Tracings - A Multimodal Approach to Music Information Retrieval

Kristian Nymoen
University of Oslo
Department of Informatics
Postboks 1080 Blindern
0316 Oslo, Norway
krisny@ifi.uio.no

Baptiste Caramiaux
IMTR Team
IRCAM, CNRS
1 Place Igor Stravinsky
75004 Paris, France
Baptiste.Caramiaux@ircam.fr

Mariusz Kozak
University of Chicago
Department of Music
1010 East 59th Street
Chicago, IL 60637, USA
mkozak@uchicago.edu

Jim Torresen
University of Oslo
Department of Informatics
Postboks 1080 Blindern
0316 Oslo, Norway
jimtoer@ifi.uio.no

## ABSTRACT

This paper investigates differences in the gestures people relate to *pitched* and *non-pitched* sounds respectively. An experiment has been carried out where participants were asked to move a rod in the air, pretending that moving it would create the sound they heard. By applying and interpreting the results from Canonical Correlation Analysis we are able to determine both simple and more complex correspondences between features of motion and features of sound in our data set. Particularly, the presence of a distinct pitch seems to influence how people relate gesture to sound. This identification of salient relationships between sounds and gestures contributes as a multi-modal approach to music information retrieval.

## Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Signal analysis, synthesis, and processing*

## Keywords

Sound Tracing, Cross-Modal Analysis, Canonical Correlation Analysis

## 1. INTRODUCTION

In recent years, numerous studies have shown that gesture, understood here as voluntary movement of the body produced toward some kind of communicative goal, is an important element of music production and perception. In the case of the former, movement is necessary in performance on acoustic instruments, and is increasingly becoming an important component in the development of new electronic musical interfaces [17]. As regards the latter, movement synchronized with sound has been found to be a universal feature of musical interactions across time and culture [14]. Research has shown both that the auditory and motor regions of the brain are connected at a neural level, and that listening to musical sounds spontaneously activates regions responsible for the planning and execution of movement, regardless of whether or not these movements are eventually carried out [4].

Altogether, this evidence points to an intimate link between sound and gesture in human perception, cognition, and behavior, and highlights that our musical behavior is inherently multimodal. To explain this connection, Godøy [6] has hypothesized the existence of *sonic-gestural objects*, or mental constructs in which auditory and motion elements are correlated in the mind of the listener. Indeed, various experiments have shown that there are correlations between sound characteristics and corresponding motion features.

Godøy et al. [7] analyzed how the morphology of sonic objects was reflected in sketches people made on a digital tablet. These sketches were referred to as *sound tracings.* In the present paper, we adopt this term and expand it to mean a recording of free-air movement imitating the perceptual qualities of a sound. The data from Godøy's experiments was analyzed qualitatively, with a focus on the causality of sound as impulsive, continuous, or iterative, and showed supporting results for the hypothesis of gestural-sonic objects.

Godøy and Jensenius [8] have suggested that body movement could serve as a link between musical score, the acoustic signal and aesthetic perspectives on music, and that body movement could be utilized in search and retrieval of music. For this to be possible, it is essential to identify pertinent motion signal descriptors and their relationship to audio signal descriptors. Several researchers have investigated motion signals in this context. Camurri et al. [1] found strong correlations with the quantity of motion when focusing on recognizing expressivity in the movement of dancers. Fur-

*MM'11,* November 28–December 1, 2011, Scottsdale, Arizona, USA.

thermore, Merer et al. [12] have studied how people labeled sounds using causal descriptors like "rotate", "move up", etc., and Eitan and Granot studied how listeners' descriptions of melodic figures in terms of how an imagined animated cartoon would move to the music [5]. Moreover, gesture features like acceleration and velocity have been shown to play an important role in synchronizing movement with sound [10]. Dimensionality reduction methods have also been applied, such as Principal Component Analysis, which was used by MacRitchie et al. to study pianists' gestures [11].

Despite ongoing efforts to explore the exact nature of the mappings between sounds and gestures, the enduring problem has been the dearth of quantitative methods for extracting relevant features from a continuous stream of audio and motion data, and correlating elements from both while avoiding *a priori* assignment of values to either one. In this paper we will expand on one such method, presented previously by the second author [2], namely the Canonical Correlation Analysis (CCA), and report on an experiment in which this method was used to find correlations between features of sound and movement. Importantly, as we will illustrate, CCA offers the possibility of a mathematical approach for selecting and analyzing perceptually salient sonic and gestural features from a continuous stream of data, and for investigating the relationship between them.

By showing the utility of this approach in an experimental setting, our long term goals are to quantitatively examine the relationship between how we listen and how we move, and to highlight the importance of this work toward a perceptually and behaviorally based multimodal approach to music information retrieval. The study presented in the present paper contributes by investigating how people move to sounds with a controlled sound corpus, with an aim to identify one or several sound-gesture mapping strategies, particularly for pitched and non-pitched sounds.

The remainder of this paper will proceed as follows. In Section 2 we will present our experimental design. Section 3 will give an overview of our analytical methods, including a more detailed description of CCA. In Sections 4 and 5 we will present the results of our analysis and a discussion of our findings, respectively. Finally, Section 6 will offer a brief conclusion and directions for future work.

## 2. EXPERIMENT

We have conducted a free air sound tracing experiment to observe how people relate motion to sound. 15 subjects (11 male and 14 female) participated in the experiment. They were recruited among students and staff at the university. 8 participants had undergone some level of musical training, 7 had not. The participants were presented with short sounds, and given the task of moving a rod in the air as if they were creating the sound that they heard. Subjects first listened to each sound two times (more if requested), then three sound tracing recordings were made to each sound using a motion capture system. The recordings were made simultaneously with sound playback after a countdown, allowing synchronization of sound and motion capture data in the analysis process.

### 2.1 Sounds

For the analysis presented in this paper, we have chosen to focus on 6 sounds that had a single, non-impulsive onset. We make our analysis with respect to the sound features *pitch*, *loudness* and *brightness*. These features are not independent from each other, but were chosen because they are related to different musical domains (melody, dynamics, and timbre, respectively); we thus suspected that even participants without much musical experience would be able to detect changes in all three variables, even if the changes occurred simultaneously. The features have also been shown to be pertinent in sound perception [13, 16]. Three of the sounds had a distinct pitch, with continuously rising or falling envelopes. The loudness envelopes of the sounds varied between a bell-shaped curve and a curve with a faster decay, and also with and without tremolo. Brightness envelopes of the sounds were varied in a similar manner.

The sounds were synthesized in Max/MSP, using subtractive synthesis in addition to amplitude and frequency modulation. The duration of the sounds were between 2 and 4 seconds. All sounds are available at the project website [1]

### 2.2 Motion Capture

A NaturalPoint Optitrack optical marker-based motion capture system was used to measure the position of one end of the rod. The system included 8 Flex V-100 cameras, operating at a rate of 100 frames per second. The rod was approximately 120 cm long and 4 cm in diameter, and weighed roughly 400 grams. It was equipped with 4 reflective markers in one end, and participants were instructed to hold the rod with both hands at the other end. The position of interest was defined as the geometric center of the markers. This position was streamed as OSC data over a gigabit ethernet connection to another computer, which recorded the data and controlled sound playback. Max/MSP was used to record motion capture data and the trigger point of the sound file into the same text file. This allowed good synchronization between motion capture data and sound data in the analysis process.

## 3. ANALYSIS METHOD

### 3.1 Data Processing

The sound files were analyzed using the MIR toolbox for Matlab by Lartillot et al.[2] We extracted feature vectors describing *loudness*, *brightness* and *pitch*. Loudness is here simplified to the RMS energy of the sound file. Brightness is calculated as the amount of spectral energy corresponding to frequencies above 1500 Hz. Pitch is calculated based on autocorrelation. As an example, sound descriptors for a pitched sound is shown in Figure 1.

The position data from the OptiTrack motion capture system contained some noise; it was therefore filtered with a sliding mean filter over 10 frames. Because of the big inertia of the rod (due to its size), the subjects did not make very abrupt or jerky motion, thus the 10 frame filter should only have the effect of removing noise.

From the position data, we calculated the vector magnitude of the 3D velocity data, and the vector magnitude of the 3D acceleration data. These features are interpreted as the velocity independent from direction, and the acceleration independent from direction, meaning the combination of tangential and normal acceleration. Furthermore, the ver-

[1] http://folk.uio.no/krisny/mirum2011

[2] http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

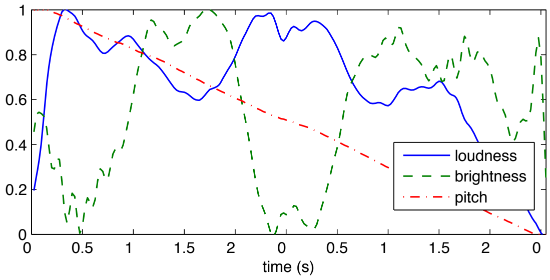

**Figure 1: Sound descriptors for a sound with falling pitch (normalized).**

tical position was used as a feature vector, since gravity and the distance to the floor act as references for axis direction and scale of this variable. The horizontal position axes, on the other hand, do not have the same type of positional reference. The subjects were not instructed in which direction to face, nor was the coordinate system of the motion capture system calibrated to have the same origin or the same direction throughout all the recording sessions, so distinguishing between the X and Y axes would be inaccurate. Hence, we calculated the mean horizontal position for each recording, and used the distance from the mean position as a one-dimensional feature describing horizontal position. All in all, this resulted in four motion features: *horizontal position*, *vertical position*, *velocity*, and *acceleration*.

## 3.2 Canonical Correlation Analysis

CCA is a common tool for investigating the linear relationships between two sets of variables in multidimensional reduction. If we let X and Y denote two datasets, CCA finds the coefficients of the linear combination of variables in X and the coefficients of the linear combination of variables from Y that are maximally correlated. The coefficients of both linear combinations are called *canonical weights* and operate as projection vectors. The projected variables are called *canonical components*. The correlation strength between canonical components is given by a correlation coefficient $\rho$. CCA operates similarly to Principal Component Analysis in the sense that it reduces the dimension of both datasets by returning $N$ canonical components for both datasets where $N$ is equal to the minimum of dimensions in X and Y. The components are usually ordered such that their respective correlation coefficient is decreasing. A more complete description of CCA can be found in [9]. A preliminary study by the second author [2] has shown its pertinent use for gesture-sound cross-modal analysis.

As presented in Section 3.1, we describe sound by three specific audio descriptors[3] and gestures by a set of four kinematic parameters. Gesture is performed synchronously to sound playback, resulting in datasets that are inherently synchronized. The goal is to apply CCA to find the linear relationships between kinematic variables and audio descriptors. If we consider uniformly sampled datastreams, and denote $\mathbf{X}$ the set of $m_1$ gesture parameters ($m_1 = 4$) and $\mathbf{Y}$ the set of $m_2$ audio descriptors ($m_2 = 3$), CCA finds two projection matrices $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_N] \in (\mathcal{R}^{m_1})^N$ and $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_N] \in (\mathcal{R}^{m_2})^N$ such that $\forall h \in 1..N$, the correlation coefficients $\rho_h = correlation(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h)$ are maximized and ordered such that $\rho_1 > \dots > \rho_N$ (where $N = \min(m_1, m_2)$).

A closer look at the projection matrices allows us to interpret the mapping. The widely used interpretation methods are either by inspecting the canonical weights, or by computing the canonical loadings. In our approach, we interpret the analysis by looking at the canonical loadings. Canonical loadings measure the contribution of the original variables in the canonical components by computing the correlation between gesture parameters $\mathbf{X}$ (or audio descriptors $\mathbf{Y}$) and its corresponding canonical components $\mathbf{XA}$ (or $\mathbf{YB}$). In other words, we compute the gesture parameter loadings $\mathbf{l}^x_{i,h} = (\text{corr}(\mathbf{x}_i, \mathbf{u}_h))$ for $1 \le i \le m_1, 1 \le h \le N$ (and similarly $\mathbf{l}^y_{i,h}$ for audio descriptors). High values in $\mathbf{l}^x_{i,h}$ or $\mathbf{l}^y_{i,h}$ indicate high correlation between realizations of the $i$-th kinematic parameter $\mathbf{x}_i$ and the $h$-th canonical component $\mathbf{u}_h$. Here we mainly focused on the first loading coefficients $h = 1, 2$ that explain most of the covariance. The corresponding $\rho_h$ is the strength of the relationship between the canonical components $\mathbf{u}_h$ and $\mathbf{v}_h$ and informs us on how relevant the interpretation of the corresponding loadings is.

The motion capture recordings in our experiment started 0.5 seconds before the sound, allowing for the capture of any preparatory motion by the subject. The CCA requires feature vectors of equal length; accordingly, the motion features were cropped to the range between when the sound started and ended, and the sound feature vectors were upsampled to the same number of samples as the motion feature vectors.

# 4. RESULTS

We will present the results from our analysis starting with looking at results from pitched sounds and then move on to the non-pitched sounds. The results from each sound tracing are displayed in the form of statistical analysis of all the results related to the two separate groups (pitched and non-pitched). In Figures 2 and 3, statistics are shown in box plots, displaying the median and the population between the first and third quartile. The rows in the plots show statistics for the first, second and third canonical component, respectively. The leftmost column displays the overall correlation strength for the particular canonical component ($\rho_h$), the middle column displays the sound feature loadings ($\mathbf{l}^y_{i,h}$), and the rightmost column displays the motion feature loadings ($\mathbf{l}^x_{i,h}$). The + marks denote examples which are considered outliers compared with the rest of the data. A high value in the leftmost column indicates that the relationship between the sound features and gesture features described by this canonical component is strong. Furthermore, high values for the sound features *loudness* (Lo), *brightness* (Br), or *pitch* (Pi), and the gesture features *horizontal position* (HP), *vertical position* (VP), *velocity* (Ve), or *acceleration* (Ac) indicates a high impact from these on the respective canonical component. This is an indication of the strength of the relationships between the sound features and motion features.

## 4.1 Pitched Sounds

The results for three sounds with distinct pitch envelopes are shown in Figure 2. In the top row, we see that the median overall correlation strength of the first canonical components is 0.994, the median canonical loading for *pitch* is

[3] As will be explained later, for non-pitched sounds we omit the *pitch* feature, leaving only two audio descriptors.

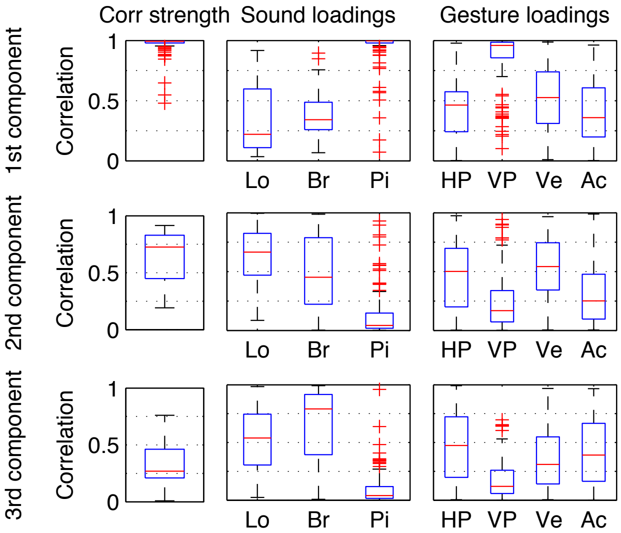

**Figure 2: Box plots of the correlation strength and canonical loadings for three pitched sounds. *Pitch* (Pi) and *vertical position* (VP) have a significantly higher impact on the first canonical component than the other parameters. This indicates a strong correlation between pitch and vertical position for pitched sounds. The remaining parameters are: *loudness* (Lo), *brightness* (Br), *horizontal position* (HP), *velocity* (Ve) and *acceleration* (Ac).**

0.997 and for *vertical position* 0.959. This indicates a strong correlation between pitch and vertical position in almost all the sound tracings for pitched sounds. The overall correlation strength for the second canonical component (middle row) is 0.726, and this canonical function suggests a certain correlation between the sound feature *loudness* and motion features *horizontal position* and *velocity*. The high variances that exist for some of the sound and motion features may be due to two factors: If some of these are indeed strong correlations, they may be less strong than the pitch-vertical position correlation For this reason, some might be pertinent to the 2nd component while others are pertinent to the 1st component. The second, and maybe the most plausible, reason for this is that these correlations may exist in some recordings while not in others. This is a natural consequence of the subjectivity in the experiment.

## 4.2 Non-pitched Sounds

Figure 3 displays the canonical loadings for three non-pitched sounds. The analysis presented in this figure was performed on the sound features loudness and brightness, disregarding pitch. With only two sound features, we are left with two canonical components. This figure shows no clear distinction between the different features, so we will need to look at this relationship in more detail to be able to find correlations between sound and motion features for these sound tracings.

For a more detailed analysis of the sounds without distinct pitch we investigated the individual sound tracings performed to non-pitched sounds. Altogether, we recorded 122 sound tracings to the non-pitched sounds; considering the first and second canonical component of these results gives a total of 244 canonical components. We wanted to analyze only the components which show a high correlation between the sound features and motion features, and for this reason we selected the subset of the components which had an overall correlation strength ($\rho$) higher than the lower quartile,[4] which in this case means a value $\leq 0.927$. This gave us a total of 99 components.

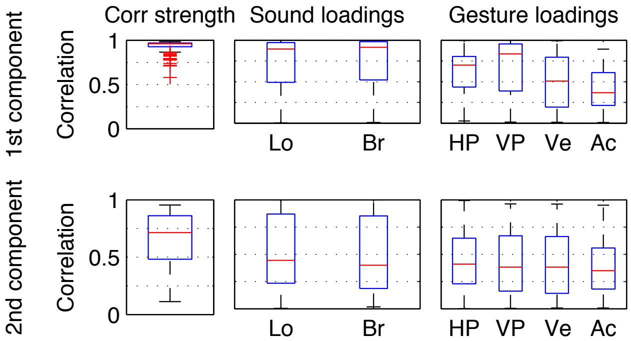

**Figure 3: Box plots of the correlation and canonical loadings for three sounds without distinct pitch.**

These 99 components all have high $\rho$-values, which signifies that they all describe some action-sound relationship well; however, since the results from Figure 3 did not show clearly which sound features they describe, we have analyzed the brightness and loudness loadings for all the recordings. As shown in Figure 4, some of these canonical components describe loudness, some describe brightness, and some describe both. We applied k-means clustering to identify the three classes which are shown by different symbols in Figure 4. Of the 99 canonical components, 32 describe loudness, 30 components describe brightness, and 37 components showed high loadings for both brightness and loudness.

Having identified the sound parameters' contribution to the canonical components, we can further inspect how the three classes of components relate to gestural features. Figure 5 shows the distribution of the gesture loadings for *horizontal position*, *vertical position* and *velocity* for the 99 canonical components. *Acceleration* has been left out of this plot, since, on average, the acceleration loading was lowest both in the first and second component for all sounds. In the upper part of the plot, we find the canonical components that are described by vertical position. The right part of the plot contains the canonical components that are described by horizontal position. Finally the color of each mark denotes the correlation to velocity ranging from black (0) to white (1). The three different symbols (triangles, squares and circles) refer to the same classes as in Figure 4.

From Figure 5 we can infer the following:

- For almost every component where the canonical loadings for both horizontal and vertical positions are high (cf. the upper right of the plot), the velocity loading is quite low (the marks are dark). This means that in the instances where horizontal and vertical position are correlated with a sound feature, velocity usually is not.

[4] The upper and lower quartiles in the figures are given by the rectangular boxes

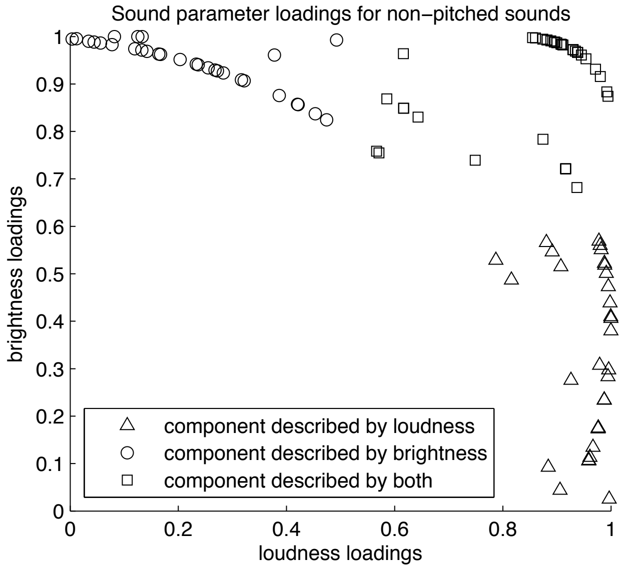

**Figure 4: Scatter plot showing the distribution of the sound feature loadings for brightness and loudness. Three distinct clusters with high coefficients for brightness, loudness, or both, are found.**

- The lower left part of the plot displays the components with low correlation between sound features and horizontal/vertical position. Most of these dots are bright, indicating that velocity is an important part in these components.
- Most of the circular marks (canonical components describing brightness) are located in the upper part of the plot, indicating that brightness is related to vertical position.
- The triangular marks (describing loudness) are distributed all over the plot, with a main focus on the right side. This suggests a tendency towards a correlation between horizontal position and loudness. What is even more interesting is that almost all the triangular dots are bright, indicating a relationship between loudness and velocity.
- The square marks (describing both loudness and brightness) are mostly distributed along the upper part of the plot. Vertical position seems to be the most relevant feature when the canonical component describes both of the sound features.

## 5. DISCUSSION

As we have shown in the previous section, there is a very strong correlation between vertical position and pitch for all the participants in our data set. This relationship was also suggested when the same data set was analyzed using a Support Vector Machine classifier [15], and corresponds well with the results previously presented by Eitan and Granot [5]. In our interpretation, there exists a one-dimensional intrinsic relationship between pitch and vertical position.

For non-pitched sounds, on the other hand, we do not find such prominent one-dimensional mappings for all subjects.

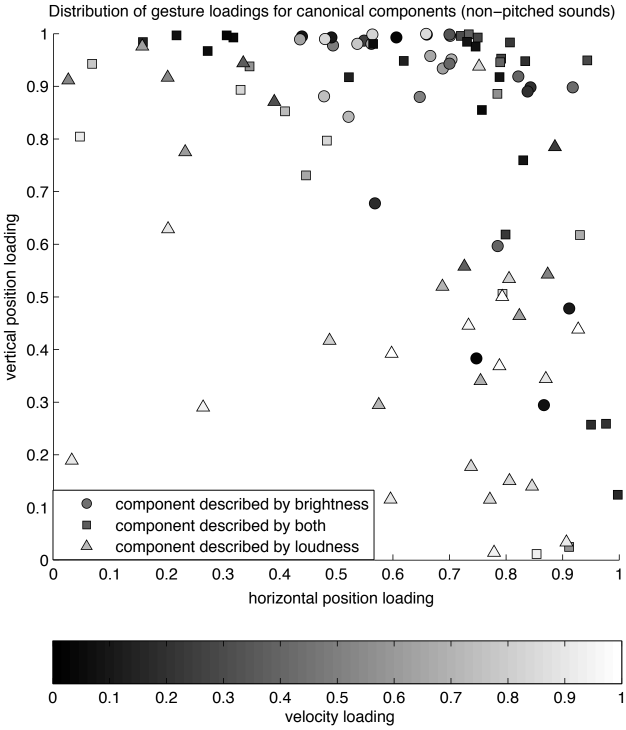

**Figure 5: Results for the 99 canonical components that had high $\rho$-values. X and Y axes show correlation for horizontal position and vertical position, respectively. Velocity correlation is shown as grayscale from black (0) to white (1). The square boxes denote components witch also are highly correlated to brightness.**

The poor discrimination between features for these sounds could be due to several factors, one of which is that there could exist non-linear relationships between the sound and the motion features that the CCA is not able to unveil. Nonlinearity is certainly plausible, since several sound features scale logarithmically. The plot in Figure 6, which shows a single sound tracing, also supports this hypothesis, wherein brightness corresponds better with the squared values of the vertical position than with the actual vertical position. We would, however, need a more sophisticated analysis method to unveil non-linear relationships between the sound features for the whole data set.

Furthermore, the scatter plot in Figure 5 shows that there are different strategies for tracing sound. In particular, there are certain clustering tendencies that might indicate that listeners select different mapping strategies. In the majority of cases we have found that loudness is described by velocity, but also quite often by the horizontal position feature. Meanwhile, brightness is often described by vertical position. In one of the sounds used in the experiment the loudness and brightness envelopes were correlated to each other. We believe that the sound tracings performed to this sound were the main contributor to the class of canonical components in Figures 4 and 5 that describe both brightness and loudness. For this class, most components are not significantly distinguished from the components that only describe brightness. The reason for this might be that peo-

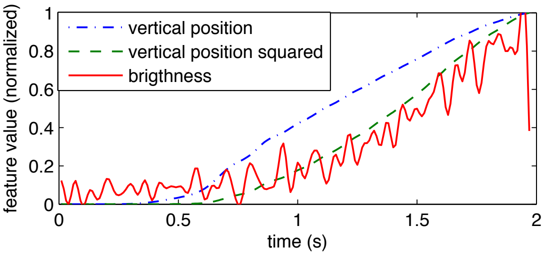

**Figure 6: Envelopes of brightness, vertical position, and vertical position squared. The squared value corresponds better with brightness than the non-squared value, suggesting a non-linear relationship.**

ple tend to follow brightness more than loudness when the two envelopes are correlated.

In future applications for music information retrieval, we envision that sound is not only described by audio descriptors, but also by lower-level gesture descriptors. We particularly believe that these descriptors will aid to extract higher-level musical features like affect and effort. We also believe that gestures will play an important role in search and retrieval of music. A simple prototype for this has already been prototyped by the second author [3]. Before more sophisticated solutions can be implemented, there is still a need for continued research on relationships between perceptual features of motion and sound.

## 6. CONCLUSIONS AND FUTURE WORK

The paper has verified and expanded the analysis results from previous work, showing a very strong correlation between pitch and vertical position. Furthermore, other, more complex relationships seem to exist between other sound and motion parameters. Our analysis suggests that there might be non-linear correspondences between these sound features and motion features. Although inter-subjective differences complicate the analysis process for these relationships, we believe some intrinsic action-sound relationships exist, and thus it is important to continue this research towards a cross-modal platform for music information retrieval.

For future directions of this research, we propose to perform this type of analysis on movement to longer segments of music. This implies a need for good segmentation methods, and possibly also methods like Dynamic Time Warping to compensate for any non-synchrony between the sound and people's movement. Furthermore, canonical loadings might be used as input to a classification algorithm, to search for clusters of strategies relating motion to sound.

## 7. REFERENCES

[1] A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1):213–225, 2003.

[2] B. Caramiaux, F. Bevilacqua, and N. Schnell. Towards a gesture-sound cross-modal analysis. In S. Kopp and I. Wachsmuth, editors, *Gesture in Embodied Communication and Human-Computer Interaction*, volume 5934 of *LNCS*, pages 158–170. Springer Berlin / Heidelberg, 2010.

[3] B. Caramiaux, F. Bevilacqua, and N. Schnell. Sound selection by gestures. In *New Interfaces for Musical Expression*, pages 329–330, Oslo, Norway, 2011.

[4] J. Chen, V. Penhune, and R. Zatorre. Moving on time: Brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Cerebral Cortex*, 18:2844–2854, 2008.

[5] Z. Eitan and R. Y. Granot. How music moves: Musical parameters and listeners' images of motion. *Music Perception*, 23(3):pp. 221–248, 2006.

[6] R. I. Godøy. Gestural-sonorous objects: Embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound*, 11(2):149–157, 2006.

[7] R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing. A preliminary study. In *2nd ConGAS Int. Symposium on Gesture Interfaces for Multimedia Systems*, Leeds, UK, 2006.

[8] R. I. Godøy and A. R. Jensenius. Body movement in music information retrieval. In *International Society for Music Information Retrieval Conference*, pages 45–50, Kobe, Japan, 2009.

[9] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson. *Multivariate Data Analysis*. Prentice Hall, New Jersey, USA, February, 2009.

[10] G. Luck and P. Toiviainen. Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis. *Music Perception*, 24(2):189–200, 2006.

[11] J. MacRitchie, B. Buch, and N. J. Bailey. Visualising musical structure through performance gesture. In *International Society for Music Information Retrieval Conference*, pages 237–242, Kobe, Japan, 2009.

[12] A. Merer, S. Ystad, R. Kronland-Martinet, and M. Aramaki. Semiotics of sounds evoking motions: Categorization and acoustic features. In R. Kronland-Martinet, S. Ystad, and K. Jensen, editors, *Computer Music Modeling and Retrieval. Sense of Sounds*, volume 4969 of *LNCS*, pages 139–158. Springer Berlin / Heidelberg, 2008.

[13] N. Misdariis, A. Minard, P. Susini, G. Lemaitre, S. McAdams, and P. Etienne. Environmental sound perception: Metadescription and modeling based on independent primary studies. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[14] B. Nettl. An ethnomusicologist contemplates universals in musical sound and musical culture. In N. Wallin, B. Merker, and S. Brown, editors, *The Origins of Music*, pages 463–472. Cambridge, Mass: MIT Press, 2000.

[15] K. Nymoen, K. Glette, S. A. Skogstad, J. Torresen, and A. R. Jensenius. Searching for cross-individual relationships between sound and movement features using an SVM classifier. In *New Interfaces for Musical Expression*, pages 259–262, Sydney, Australia, 2010.

[16] P. Susini, S. McAdams, S. Winsberg, I. Perry, S. Vieillard, and X. Rodet. Characterizing the sound quality of air-conditioning noise. *Applied Acoustics*, 65(8):763–790, 2004.

[17] D. van Nort. Instrumental listening: Sonic gesture as design principle. *Organised Sound*, 14(02):177–187, 2009.