

Analysing Correspondence Between Sound Objects and Body Motion

Kristian Nymoen, Rolf Inge Godøy and Alexander Refsum Jensenius, Jim Torresen

ACM Transactions on Applied Perception (to appear)

Submitted 4 June 2012 / Accepted 2 December 2012

Abstract

Links between music and body motion can be studied through experiments called *sound-tracing*. One of the main challenges in such research is to develop robust analysis techniques that are able to deal with the multidimensional data that musical sound and body motion present. The paper evaluates four different analysis methods applied to an experiment in which participants moved their hands following perceptual features of short sound objects. Motion capture data has been analysed and correlated with a set of quantitative sound features using four different methods: (a) a pattern recognition classifier, (b) *t*-tests, (c) Spearman's ρ correlation, and (d) canonical correlation. The paper shows how the analysis methods complement each other, and that applying several analysis techniques to the same data set can broaden the knowledge gained from the experiment.

1 Introduction

Body motion is an integral part of the experience of music. On the one hand, body motion can create sound, such as a sound-producing action on a musical instrument. As humans, we all have some ecological knowledge of how sound-producing actions relate to sound (Clarke, 2005). For instance, most of us immediately know what sound to expect when we see two cymbals that are about to crash. On the other hand, sound can lead to body motion. We experience this when music makes us tap a foot or bob our head, and music can influence basic human motion patterns, such as walking (Styns et al., 2007). As such, Leman (2008) has suggested that the human body functions as a “mediator” between physical sound waves and the mind, structuring our perception of musical sound. Music-related motion (like dance) then becomes a type of corporeal articulation of our cognition of music. Further, it has been suggested that mental images of motion play an essential role in our involvement with music, and that most features in a musical soundscape could potentially afford gestural responses by listeners (Godøy, 2006, 2010a).

It has been shown that our perception of phenomena is often done through a combination of modalities — an effect known as *cross-modality* (Stein and Meredith, 1993; Vroomen and de Gedler, 2000). While research on cross-modality to a large extent has been concerned with interactions between vision, audition and tactile perception, motion sensations have also been suggested as important to our perception — for instance in Liberman and Mattingly's (1985) *motor theory of speech perception*. This theory claims that speech is understood not only as sound, but also through our own experience with gestures that produce phonemes. Although originally conceived as a special phenomenon for speech perception, Galantucci et al. (2006) have presented convincing arguments for why this theory should be extended to perception in general.

The evidence for an interaction between motor sensations and other modalities was strengthened with the discovery of *mirror neurons* in the premotor cortex of macaques (Pellegrino et al., 1992). These neurons were shown to activate not only when the macaque performed an action, but also when seeing the action performed by someone else, and even when only hearing the sound of the action (Kohler et al., 2002).

Based on considerations like the ones presented above, it has been suggested that research on relationships between sound and body motion can help us learn more about processes in the brain (Zatorre, 2005). Furthermore, this research can contribute to more efficient solutions for so-called *query-by-gesture* in music information retrieval databases (Leman, 2008; Godøy and Jensenius, 2009). Moreover, research on correspondences between musical sound and body motion can facilitate development of new musical interfaces, where body motion is the main input source (van Nort, 2009; Nymoen et al., 2011b).

One of our contributions to research on sound and motion has been studies of so-called *sound-tracing* (Godøy et al., 2006; Nymoen et al., 2010, 2011a, 2012). The point of sound-tracing is to observe and collect data of listener’s spontaneous gestural renderings of sound, i.e. to capture listeners’ immediate associations of sound and motion. In these studies, participants have been asked to render the perceptual qualities of short sound objects through bodily motion, either in two dimensions by drawing with a pen on a digital tablet, or three dimensions by free-air hand gestures. The underlying assumption and motivation of sound-tracing experiments is that they may reveal salient sound-motion links in perception. A main challenge with this type of research is to develop techniques for capturing, storing, analysing, and visualising both action and sound in a coherent and consistent manner. This paper addresses the analysis of such data, evaluating strengths and weaknesses of using standard correlation measures, hypothesis tests, and pattern recognition classifiers as analysis tools.

2 Background and Motivation

Both music and body motion are highly multidimensional phenomena, and research on how people move to musical sound entails several challenges. Approaches for studying people’s bodily response to music can be ordered along a continuum, where one extreme involves looking at unconstrained full body motion. This has the advantage of being the most “natural” way of moving to music, allowing people to move as they please. On the downside, the multidimensionality of unconstrained full-body motion necessitates use of high-level features or dimensionality reduction techniques which do not provide precise details about the body motion, see e.g. (Camurri et al., 2003; Burger et al., 2012). A radically different approach to studying music-related body motion is by constraining the response down to a one-dimensional space, for instance by instructing people to move a slider or push a button. In this way, a controlled and easily measurable response is obtained, but such a reduction of body motion has limitations in a musical context. A similar argument can be made for the sound stimuli used, with a continuum spanning from a highly multidimensional musical piece, down to a simple sine tone or noise burst which can be described precisely with low-level features such as frequency or duration. In summary, research on music and body motion involves a trade-off between doing controlled experiments and using ecologically valid stimuli and response data.

Our focus of research is on body motion in a musical context. And to be able to relate the auditory stimuli to music, we have chosen to use sound files which are slightly more complex than sine tones and noise bursts in our experiment (these will be presented more in detail in Section 3.2). We adapt ideas from the French composer and theorist Pierre Schaeffer (1966), who in his work on aesthetics for *musique concrète* presented the *sonic object* as a basic unit that can be put into a musical context and thus made into a *musical object*. Sonic objects are typically 0.5–5 s in duration, and they are perceived holistically. Interestingly, Schaeffer’s findings correspond well with later research on *auditory objects* (Bregman, 1990) and *stream segregation* (van Noorden, 1975). Schaeffer introduced a typology for sonic objects which included three categories based on the overall dynamic envelope of the object. First, the *impulsive* type refers to a sharp attack followed

by a decaying envelope, such as the sound from plucking a string. Second, the *sustained* type refers to a more or less steady dynamic envelope, such as a trumpet sound. Third, the *iterative* type refers to rapidly repeated impulsive onsets that fuse together, such as in a harp glissando. Rolf Inge Godøy (2006) pointed out an interesting link between Schaeffer’s typology and the notion of *motor mimesis*, meaning the visualisation of gestural trajectories and sound-producing actions as part of music cognition: Schaeffer’s three types of dynamic envelopes are in fact closely connected to sound-producing actions. An impulsive sound relates to an impulsive energy transfer from, e.g., a mallet to a membrane. Next, a sustained sound relates to a continuous transfer of energy, such as bowing a string. And lastly, an iterative sound is produced by rapidly repeated back-and-forth energy transfers, for instance like the repeated excitations of the membrane in a drum roll.

Godøy (2006) further emphasised how each of the three types of dynamic envelopes can be found for very different sounds made by different sources, and that each of the corresponding types of sound-producing actions can be performed with different effectors, demonstrating what is called *motor equivalence* in motor control literature (Rosenbaum, 2001). As such, both sound objects and sound-producing actions seem to have a degree of abstraction, some higher level where they possibly share certain features and are perceived as similar. This abstraction of sound objects and sound-producing actions is our main interest when studying music-related body motion, and the larger purpose in our research is to study how generic features of sound and motion can reveal perceptual similarities between different sounds, between different actions, and between sounds and actions.

2.1 Describing Sound and Motion

As humans we describe sound and motion at a higher level than the raw sound file and the unprocessed motion capture data. These descriptions, or *features*, can concern an entire sound object or action object, such as labelling a sound-producing action as “impulsive”. While this label gives certain associations about the action, it does not provide detailed information about the action itself. Even if we have two actions of the same type, let us say hand claps, these may be different with respect to the trajectories of the hands, the velocities, angles, etc. Since we are also interested in distinguishing between such nuances, we need lower-level features that describe continuous changes in velocity, position, etc. For now, we will focus on conceptual differences between descriptors, and we will return more in detail to the features we have used in our analysis in Section 3.

2.1.1 Timescales

Schaeffer’s typological classification of sonic objects is concerned with the overall dynamic and pitch envelopes of the object (Schaeffer, 1966). His theory also included a description of internal features of the sonic objects. First, “allure”, which may be translated into “motion”, describing slower fluctuations of harmonic content, pitch, and loudness, and second, “grain”, which relates to a perceived graininess of the sound. Schaeffer’s morphological descriptions of sonic objects derived from experiments on 78 RPM discs and magnetic tape may seem crude compared to the precise descriptions that have been made possible with digital technology, e.g. (Vassilakis, 2001; Peeters et al., 2011). However, his distinction between descriptions of the entire sonic object, and descriptions of internal features of the object is important.

Godøy (2010b) coined the terms *micro*, *meso* and *macro* levels of *timescales* for sonic objects which can help us understand how objects can be described from different timescale perspectives. The micro level describes continuously varying features, at a timescale shorter than the entire object. For a sonic object, this may refer to the pitch at a certain time, and for an action object, the velocity at a certain time. Next, the meso level describes an entire object, for instance like Schaeffer’s typological description of onset type. For body motion, a label such as two-handed or one-handed could describe an action at this level. The macro timescale level refers to sequences of sonic objects, and thus features operating at the macro level describe events like melodic phrases or some development within larger sections of music.

In this paper, we are mostly interested in the micro and meso timescales, and the analysis methods we use apply to these. Just like features of motion or sound may describe an entire object or a continuously varying feature within the object, the relationship between a sound object and a corresponding motion recording can be analysed on a very small (micro) level, or at a chunk (meso) level. We shall use the terms *global features* and *time-varying features* (adopted from (Peeters et al., 2011)) to distinguish between the two levels. In our experimental setup (presented in Section 3), the raw data from the motion capture system represents sampling of spatial positions of the body at regular time intervals, and time-varying features describe some aspect of the sound-tracing at each timeframe. Global features are typically based on descriptive statistics of the time-varying features, e.g. mean or maximum, and correspond to the meso level.

2.2 Methods for Analysing Sound-Tracings

Several experiments within systematic musicology in the last decades have observed how people relate motion or shapes to short sound objects. Some of these have collected participant’s responses as visual or verbal metaphors, e.g. (Walker, 1987, 2000; Eitan and Granot, 2006; Pirhonen, 2007; Merer et al., 2008), while others have used sensing technologies to capture the (more or less unconstrained) motion of participants, e.g. (Haga, 2008; Caramiaux et al., 2010; Kozak et al., 2011; Kussner, 2012). Experiments with responses of the former type can typically perform statistical analyses directly on the response data, e.g. by comparing the number of participants that provided one response to the number of participants with another response. Also, when dealing with global features calculated from continuous response data, a sufficient form of analysis can be to apply appropriate statistical tests and evaluate the statistical results in light of the experiment conditions. Experiments by Repp (1995), Camurri et al. (2003), and Leman et al. (2004) are examples of this.

When the research objective involves studying people’s preferred use of a motion-based musical interface, experiments using continuous response data may be more appropriate, as this type of response would be closer to the actual use scenario. The use of time series in analysis, and particularly multidimensional time series like time-varying motion features, presents several challenges. A common practice in music cognition research has been to use *Pearson product-moment correlation* to assess similarities between time-varying sound features and response features, such as perceived “tension” or “emotion” in music. Emery Schubert (2002) criticised this practice, arguing that the serial nature of the data violates an important principle for significance testing, namely the assumption of independent, normally distributed data.

Schubert suggested that the problem can be reduced by applying the non-parametric measure *Spearman’s ρ* instead of the parametric Pearson correlation, and also using first or second order differentiations of the data. Further, Schubert admits that researchers can report with confidence the ranking of coefficients within a data set, but that “problems arise when (1) the correlation coefficient is compared with coefficients from other sources, and (2) when the significance of the coefficient is taken ‘literally’ ” (Schubert, 2002, p. 219).

In addition to Schubert’s own research, Vines et al. (2006) and Kussner (2012) followed Schubert’s suggestion of applying Spearman’s ρ correlation when comparing time series of sound data to other time series. A different approach was taken by Caramiaux et al. (2010), who applied canonical correlation analysis to analyse the relationship between features of sound and motion in musical performance. We shall compare the usefulness of several of these techniques, using the dataset collected in a single experiment.

3 Experimental setup

The data set that has been analysed in this paper was recorded in autumn 2010. This section will present the experimental setup, as well as the sound and motion features that have been used in the analysis process.

3.1 Participants, Task, and Equipment

We recruited 38 people to the experiment — 29 male and 9 female. The participants were told to imagine that they could create sound by moving their hands in the air. They were presented with short sound objects, and instructed to move their hands as if they were creating the sound themselves. The hand motion was captured with a Qualisys motion capture (mocap) system, with 9 Oqus 300 cameras, sampling at 100 Hz.

18 monophonic sound files (described more in detail below) were presented to the participants, played through two Genelec 1031A speakers, positioned approximately 3 m to each side of the participant. Each participant started by listening to all the sounds played one by one in succession. Then, one sound at a time was played back in a random order. For each sound the participant listened to the sound once without moving, and then a second time while their motion was recorded. Each time, the sound playback was preceded by a countdown, to allow participants to anticipate the beginning of the sound playback. The mocap recording for each sound file started 0.5 seconds before the sound playback, and ended at the end of the sound file.

After the recording session, the participants were asked to comment on the experiment and rate their own level of musical training as (1) extensive, (2) medium, or (3) little/no training on a questionnaire.

3.2 Sound files

18 sound files were designed in the visual programming environment Max5¹ using a combination of *frequency modulation* (FM) synthesis and subtractive synthesis. The duration of each sound file was 3 seconds. Godøy (2010a) has suggested several features of musical sound which could afford gestural responses, among them *pitch contours*, *timbral contours*, and *dynamic contours*. In order to keep the number of variables within a manageable range, our experiment is concerned with these three sound feature types. The reason for not considering the spatial content of sound was to limit the number of variables in the stimuli. We suspect that adding the spatial dimension of sound to the experiment would have inspired subjects to follow the perceived location of the sound. This is an interesting aspect which we plan to pursue further in future experiments.




Pitch was manipulated by changing the carrier frequency of the FM synthesiser, and non-pitched sounds were made by replacing the FM synthesis with pink noise. *Timbre* is a multidimensional sound feature that is difficult to describe precisely (Grey, 1977; McAdams et al., 1995), and so we chose to manipulate only the *brightness* of the sound, since this is one of the dimensions that distinguishes the strongest between sounds (Wessel, 1979). The brightness was manipulated by changing the centre frequency of a bandpass filter. Finally, the dynamic contour was manipulated by multiplying each sound file with one out of two envelope functions which were (1) slowly increasing and decreasing (non-impulsive), and (2) quickly increasing and slowly decreasing (impulsive).

We have used the MIR toolbox for Matlab (Lartillot et al., 2008) to extract time-varying feature vectors of the sounds. The *dynamic envelope* or *loudness envelope*, is here simplified and calculated by the RMS value of the audio waveform. Brightness can be measured in various ways, and we have chosen to represent it by the *spectral centroid*, denoting the barycentre of the frequency spectrum (Wessel, 1979). The MIR toolbox uses an autocorrelation function to calculate pitch. Since the sound files were designed from scratch, the pitch vectors were easily verified by comparing against the values used for the FM carrier frequency. Table 1 summarises the sound features we will use in the paper.

In order to be able to evaluate the significance of e.g. a steady spectral centroid versus a rising spectral centroid under various conditions, a parametric tree of features was used for twelve of the sound files (Figure 1). The set of stimuli also included two sound files where two peak filters were used instead of the bandpass filter. In Sound 13, the peak frequencies of the two filters moved toward each other, and in Sound 14, away from each other. This causes a distinctly different timbral

¹<http://www.cycling74.com>

Table 1: Sound features used in the analysis. The symbols will be used in illustrations throughout the paper.

Perceptual	Symbol	Calculated by
Loudness		The RMS value of the sound.
Brightness		The spectral centroid of the sound.
Pitch		Autocorrelation function in the MIR toolbox (<i>mirpitch</i>)

unfolding than in the other sounds. Sounds 13 and 14 were included to be able to infer whether such formant changes in sound would lead to different motion responses. Still, for the sake of consistency, Sounds 13 and 14 were evaluated by the same features as the rest of the data set. A non-impulsive envelope function was applied to Sounds 1–14.² Next, the impulsive envelope function was applied to some of the sound files, providing 4 sounds with a sharp attack and a protracted decay.

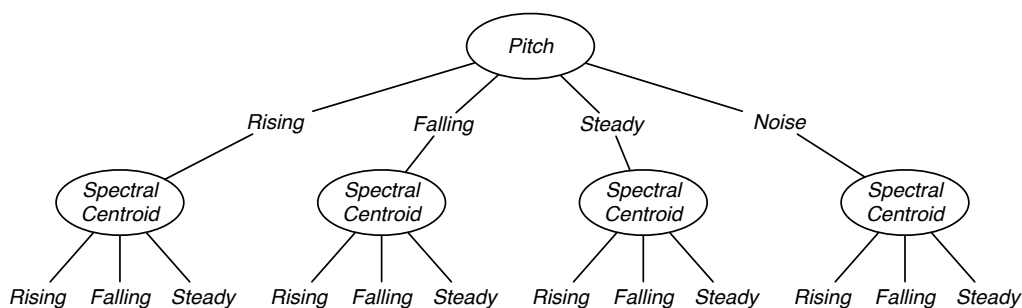


Figure 1: Sounds 1–12 were designed by a parametric tree of features.

The different types of sounds allow analysis of responses to *impulsive* versus *non-impulsive* sound objects, as well as *pitched* versus *non-pitched*. Furthermore, we have defined a category for “rising” sound objects containing sounds with either rising pitch, rising spectral centroid or both, and similarly “falling” sound objects with falling pitch or spectral centroid. Sounds with steady spectral centroid in combination with a steady pitch or no perceivable pitch are referred to as “steady” sounds. An overview of the sound files used in the experiment is given in Table 2, and the sound files are available for download and further analysis.³

Table 2: Description of the sound files used in the experiment. The columns display the pitch envelope, spectral centroid envelope and the dynamic envelope of each sound file. The numbers 1 and 2 for non-impulsive dynamic envelopes refer to (1) a bell-shaped envelope, and (2) a continuously increasing envelope as explained above.

Sound	Pitch	Sp.Centroid	Dyn.Env.	Sound	Pitch	Sp.Centroid	Dyn.Env.
1	Rising	Falling	Non-impulsive ¹	10	Noise	Falling	Non-impulsive ¹
2	Falling	Rising	Non-impulsive ¹	11	Noise	Rising	Non-impulsive ²
3	Falling	Falling	Non-impulsive ¹	12	Noise	Steady	Non-impulsive ²
4	Rising	Rising	Non-impulsive ¹	13	Steady	Rising slightly	Non-impulsive ²
5	Rising	Steady	Non-impulsive ²	14	Steady	Falling slightly	Non-impulsive ²
6	Falling	Steady	Non-impulsive ²	15	Rising	Falling	Impulsive
7	Steady	Falling	Non-impulsive ¹	16	Steady	Steady	Impulsive
8	Steady	Rising	Non-impulsive ¹	17	Noise	Steady	Impulsive
9	Steady	Steady	Non-impulsive ²	18	Noise	Falling	Impulsive

²The combination of pitch envelope and the centre frequency envelope of the bandpass filter resulted in two different dynamic envelopes for the non-impulsive sounds: *bell-shaped* and *increasing*. Our analysis of dynamic envelopes is mainly concerned with the distinction between impulsive and non-impulsive. The two different envelopes are marked with numbers in Table 2.

³<http://fourms.uio.no/downloads/audio/sound-tracings/soundsST2.zip>

3.3 Motion Capture Data

3.3.1 Pre-processing





Optical infrared marker-based motion capture technology is sensitive to occlusion of the reflective markers. Whenever this happens, there will be gaps in the recorded data. Short gaps in our recordings were gap-filled by interpolating with a spline function, and recordings with long gaps were discarded. Some participants had a large number of their recordings discarded, due to poor system calibration in some sessions, and also because some participants repeatedly moved their hands outside the capture space. For participants with more than 1/3 (i.e. more than six) of the recordings being discarded, the participant data was left out of the analysis. Of the 684 recordings, 156 were removed for the reasons mentioned above. The position data was smoothed with a 5 samples running mean smoothing filter. A similar smoothing was performed for each derivative level in the feature extraction process.

3.3.2 Motion Features

In order to compare the motion data to the sound features, we have calculated various features from the mocap data. Since the design of musical instruments is one of the underlying purposes for this research, we consider time-varying features that could be calculated in realtime with a reasonably low amount of latency. Thus, the features have been based on the position data, with first and second order differentiations, filtering, and the distance between the hands.

The mocap data describes the position of the participant’s hands in relation to a fixed coordinate system in the room. The participants in our experiment were not instructed specifically where they should stand, or in which direction to face. Thus, horizontal position in relation to the room was not used in our analysis. Vertical position, on the other hand, is the distance to the floor, and since the floor level and the direction of gravity were equal references for all participants, the vertical position was used as a motion feature. To compensate for differences in height between participants, the the data series of each participant were normalised based on their overall maximum and minimum. The time-varying features that were used in the analysis process are summarised in Table 3. Additionally, 11 global features were calculated, summarised in Table 4. The first 7 features are calculations of mean and standard deviation (SD) of the time-varying feature vectors. The following 4 features required more processing, hence these are explained in more detail.

Table 3: The time-varying motion features that were used in the analysis

Feature	Symbol	Calculated by
Vertical position		Distance to the floor
Absolute velocity		Euclidean distance between successive position samples
Absolute acceleration		Euclidean distance between successive samples of the first derivative of position
Hand distance		Euclidean distance between the hands

4 Analysis I: Pattern Classification

In our first analysis, a classifier was used to find clusters in the sound-tracings. The aim was to see if a computer could identify the sounds, based on the motion features of each sound-tracing. A classification result better than chance would suggest that similarities between the hand motion from different participants exist. We have analysed the classification results to assess similarity between motions related to different sounds.

Table 4: The global motion features that were used in the analysis. These were normalised per subject.

Feature	Description
VerticalVelocityMean	The mean value of the first derivative of <i>vertical position</i> .
AbsVelocityMean	The mean value of <i>absolute velocity</i> .
AbsVelocitySD	The SD of <i>absolute velocity</i> .
AbsAccelerationMean	The mean value of <i>absolute acceleration</i> .
HandDistanceMean	The mean value of <i>hand distance</i> .
HandDivergenceMean	The mean value of the first derivative of <i>hand distance</i> .
HandDivergenceSD	The SD of the first derivative of <i>hand distance</i> . Together with <i>HandDivergenceMean</i> , this feature explains how steady the hands' tendency of moving apart or towards each other is.
VerticalEnergyUM	The sum of the signed vertical kinetic energy of a unit mass. Or to be more precise: <div style="text-align: center; margin: 10px 0;"> $VerticalEnergyUM = \sum_{k=1}^n v_k \cdot v_k$ </div> where v_k is the k -th sample of the vertical velocity vector of length n . This differs from the mean vertical position by letting faster motion count more than slower motion. For instance, fast upward motion followed by a slow downward motion back to the starting position would give a positive value, whereas the <i>VerticalVelocityMean</i> feature would be 0.
Symmetry	This feature describes the difference between the left and right hand in terms of vertical position and horizontal velocity. Vertical position should be similar for the two hands in order to be symmetric. Horizontal position, on the other hand, would only be similar if the hands were held directly in front of the body, so <i>velocity</i> is a better measure than <i>position</i> for horizontal symmetry. <div style="text-align: center; margin: 10px 0;"> $Symmetry = \sum_{k=1}^n \left v_k^{\text{leftHorizontal}} - v_k^{\text{rightHorizontal}} \right + \left p_k^{\text{leftVertical}} - p_k^{\text{rightVertical}} \right$ </div> where $p_k^{\text{leftVertical}}$ and $p_k^{\text{rightVertical}}$ denote the k -th sample of the vertical position vector of length n of the left and right hand, respectively. Similarly $v_k^{\text{leftHorizontal}}$ and $v_k^{\text{rightHorizontal}}$ denote the k -th sample of the horizontal velocities of the two hands.
OnsetAcceleration	The mean acceleration in the interval from 20 samples before to 50 samples after the sound started.
Shaking	The mean value of the <i>absolute acceleration</i> time-varying feature filtered by a 30 samples running mean filter. The effect of this is that smaller spikes or impulsive actions are filtered out, but high acceleration due to continuous shaking is still measured.

A pattern classifier can be trained to recognize certain patterns in a data set, and predict the classes of unknown instances. To train and validate the performance of a classifier, we typically use a *training set* and a *validation set*. The former consists of a number of instances with attribute vectors (features) and class labels that are used to train the classifier. The validation set is kept separate from the training set, and subjected to classification after the training is complete. The classifier performance is evaluated by comparing the true classes of the instances in the validation set to the classes predicted by the classifier (Duda et al., 2000).

We have used a Support Vector Machine (SVM) classifier to classify the data, as has also been used in previous work (Nymoen et al., 2010). The 11 global motion features from each sound-tracing were used as input to the classifier, paired with the sound number as the *class*. In this way, we trained the classifier to recognize the sound object based on the motion that was performed to it. We used RapidMiner to design the classifier (Mierswa et al., 2006), and *cross-validation* (leave-one-out) was used for validation. This means that the classifier was trained several times, and each time, one sound-tracing was left out of the training set and used for validation, as shown in Figure 2 (Duda et al., 2000).

Table 5 shows the *confusion matrix* from our experiment. This displays the distribution of the predictions made by the classifier for each instance and the true class of the instance. The 18 sounds used in the experiment give 18 classes. Each column (T4, T5, etc., T = true) contains the sound-tracings that belong to specific classes, and each row (P4, P5, etc., P = predicted) contains the classification made by the classifier. For instance, for class 8 (the T8 column), the classifier classified two instances as class 4 (the P4 row), three as class 5 (the P5 row), five were correctly

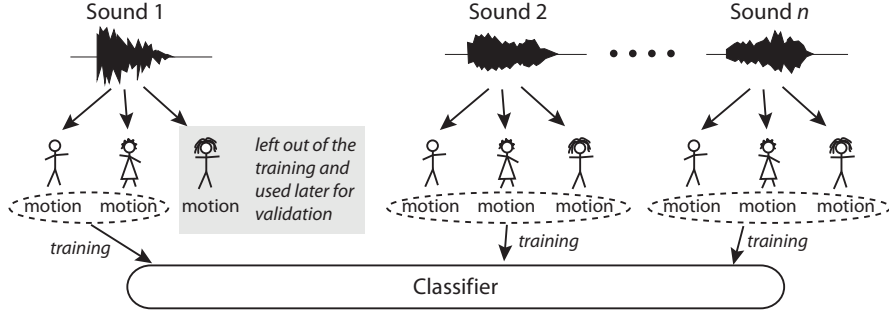


Figure 2: The classifier was trained based on the motion features, using the sound number as the class label. One motion was left out of the training data, and the classifier predicted which class this belonged to. The process was repeated such that all of the motion recordings were left out once.

Table 5: Confusion matrix showing the predictions made by the classifier (P) and the true classes of the instances (T). The table has been sorted such that sounds with similar labels are placed together. In addition to the Class Recall (CR) and Class Precision (CP) for each sound, the Group CR/CP show the corresponding value when we choose to let prediction count as correct as long as it is within the same group as the true class (the same grey area).

	← rising →				rising/falling			← falling →				← steady →				impulsive		CP	Groups	
	T4	T5	T8	T11	T1	T15	T2	T3	T6	T7	T10	T12	T13	T14	T9	T16	T17	T18		
P4	3	7	2	3	2	1	0	0	0	0	0	1	0	0	0	0	0	1	0.15	0.69
P5	2	2	3	3	9	1	0	0	0	0	0	1	0	3	1	0	0	0	0.08	
P8	3	6	5	3	2	5	0	0	0	1	0	4*	4*	2*	3*	0	0	0	0.13	
P11	7	1	1	12	1	3	0	0	0	1	3	5	0	1	1	0	0	1	0.32	
P1	9	7	3	2	8	6	1	1	1	3	1	2	3	3	1	0	0	0	0.16	
P15	2	0	0	1	2	1	1	0	2	2	0	0	0	0	0	1	1	1	0.07	0.65
P2	0	0	0	1	1	1	1	2	5	0	1	0	0	0	0	0	0	0	0.08	
P3	0	0	0	1	0	3	6	9	5	3	4	0	0	0	0	0	1	6	0.24	
P6	0	0	0	0	1	3	6	7	6	2	7	0	1	2	0	2	0	0	0.16	
P7	1	0	4	0	1	3	5	4	3	7	3	3*	3*	3*	2*	2	1	1	0.15	
P10	0	0	1	2	1	0	2	4	3	0	8	3	1	1	0	0	0	0	0.31	0.60
P12	0	0	4*	1	0	0	1	0	0	3*	0	5	4	2	3	0	0	0	0.22	
P13	0	0	2*	1	0	0	1	0	1	2*	1	1	3	7	1	2**	1	0	0.13	
P14	0	0	1*	0	0	0	0	1	0	1*	0	2	1	0	4	2**	0	0	0.00	
P9	0	2	5*	0	1	0	0	0	1	4*	1	3	9	3	14	2**	0	0	0.31	
P16	0	0	0	0	0	0	1	0	0	2	0	0	1	3	0	12	6	7	0.38	0.83
P17	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	10	7	0.50	
P18	0	0	0	0	0	2	1	1	2	0	0	0	0	0	0	5	11	7	0.24	
CR	0.11	0.08	0.16	0.40	0.28	0.03	0.04	0.31	0.21	0.23	0.28	0.17	0.10	0.00	0.47	0.40	0.32	0.23		
Groups	0.74						0.72						0.52				0.73			

classified as class 8 (the P8 row), and so forth. The correctly classified instances are found along the diagonal, and are shown in bold face. *Class precision* (CP) and *class recall* (CR) are displayed in the rightmost column and the lower row, respectively, and are given by:

$$CP_i = \frac{||R_i \cap A_i||}{||A_i||} \quad \text{and} \quad CR_i = \frac{||R_i \cap A_i||}{||R_i||},$$

where $||A_i||$ denotes the number of instances classified as i , and $||R_i||$ denotes the total numbers of instances in class i . In other words, CP denotes the correctly classified instances of a class i divided by the number of instances classified as class i , and CR denotes the correctly classified instances of a class i divided by the true number of instances in class i . Together, CP and CR denote the accuracy and precision of the classification.

At first glance, the performance of the classifier is not very impressive. Although certain classes have CR and CP values above 30 %, the majority of the classes have CP or CR values of less than 20 %. On further inspection, we observe that the misclassification of instances is somewhat

systematic, and that sounds are more often mistaken for sounds with similar sound features than for sounds with distinctly different sound features. The rows and columns in the confusion matrix in Table 5 have been sorted to display this, and the grey fields are used to denote sounds with similar sound features. The first four rows (and columns) contain classes where the sound can be classified as “rising”. Sound 4 had a rising pitch, and rising spectral centroid, Sound 5 had rising pitch and steady spectral centroid, Sound 8 had a stable pitch and rising spectral centroid, and Sound 11 was based on noise and had a rising spectral centroid (cf. Table 2). Rows 5–7, denoted by rising/falling, contain sounds where the envelopes of pitch and spectral centroid move in opposite directions. Rows 8–11 contain “falling” sounds (as opposed to the “rising” sounds in rows 1–4). Rows 12–15 contain sounds classified as “stable”, meaning no distinct rising or falling tendency in pitch or spectral centroid. Rows 16–18 contain impulsive sounds (all other sounds apart from Sound 15 had a non-impulsive onset).

When we read the table in light of these classes, we can make some assumption for why a particular sound-tracing was misclassified. For instance, for Sound 5, only two sound-tracings were classified correctly. However, 21 of the sound-tracings were classified as Sounds 1, 4, 8 or 11, which all have some rising tendency, similarly to Sound 5. For this extended class, the CR is 74 % and CP is 69 % as shown in the table.

It is interesting that classes 7 and 8 are often mistaken for steady sounds and vice versa (marked by * in Table 5). Both of these sound files had a stable pitch, and varying spectral centroid. The misclassification indicates that the sound-tracings to these sound files have equally much in common with the steady sound-tracings, as with the rising and falling classes, which again suggests that some subjects mainly focused on the steady pitch, while others traced the envelope of the spectral centroid.

The table also shows that the impulsive sounds were more often mistaken for sounds labelled “falling” than for sounds labelled “rising”. This may be because participants would lift their hands to make an accentuated attack, and then lower them and come to a rest right after the attack was made. Also notice how pitch influenced the impulsive sounds; six out of the 30 sound-tracings to the pitched, impulsive sound were mistaken for steady, non-impulsive sounds (marked by ** in Table 5).

A similar classification was performed using only expert and only non-expert subjects (groups 1 and 3 from Section 3.1). These classifications gave slightly lower classification accuracies, most likely due to a smaller training set. However, the results were not much different than when the whole data set was used.

In summary, the pattern classification method does identify clusters in the sound-tracings, and verifies that sound-tracings based on similar sounds are similar to each other. However, the method does not provide clear answers to which motion features are most pertinent to the classifier.

5 Analysis II: Hypothesis Testing

The results of the pattern recognition analysis in the previous section indicated certain similarities between sound-tracings within various groups of sounds. However, the classifier did not provide information on the pertinence of the individual motion features to the classification results. In a previous publication, we applied standard statistical tests to the same data set based on a set of assumptions of how certain motion features would distinguish well between different groups of sounds (Nymoen et al., 2012). In this section, we will briefly summarise this method and the findings from the other publication.

Various groups of sound-tracings were compared to each other. We provide as an example the comparison between sound-tracings of “rising” sound objects versus “falling” sound objects. Visualisations of the data seemed to verify our assumption that “rising” sounds would show higher values for VerticalVelocityMean than “falling” sounds, and so a *t*-test was performed with a null-hypothesis stating the opposite case. Next, the same was also tested for only the musical expert participants, and then for only musical non-experts. Finally, the experts’ responses to “rising”

sounds were compared with the non-expert responses to the same sounds, and then the responses to “falling” sounds was compared between experts and non-experts.

All in all, 20 t -tests were performed, meaning that the experiment-wise error rate is potentially high. However, this error rate was not controlled in the previous study. After correcting for repeated measurements for the present publication, using the Holm-Bonferroni method (Holm, 1979), nine of the findings were found to be significant at $\alpha = 0.05$, including the three initial hypotheses, listed in Table 6. The same tests applied to only expert participants and to only non-experts also showed significant results.

Table 6: Results from unpaired t -tests, showing comparisons between (1) the OnsetAcceleration levels of impulsive versus non-impulsive sounds, (2) the VerticalVelocityMean of rising versus falling sounds, and (3) the AbsAccelerationMean of pitched versus noise sounds. df denotes the degrees of freedom (i.e. the size of the test material), t is the t-value related to the t -test, and p is the probability that the two data sets are equally distributed.

Motion feature	Comparison	df	t	p
OnsetAcceleration	Impulsive vs non-impulsive sounds	526	13.65	< 0.01
VerticalVelocityMean	Rising vs falling sounds	284	18.89	< 0.01
AbsAccelerationMean	Pitched vs noise-based sounds	179	5.53	< 0.01

The t -tests revealed that the onset characteristics of the sounds influenced OnsetAcceleration in the sound-tracings. Furthermore, the difference between average vertical velocity for falling and rising sounds was highly significant, with an average upward tendency for rising sounds, and a downward tendency for falling sounds. Finally, the mean acceleration was significantly higher for noise-based sounds than for pitched sounds. Additionally, not shown in the table, expert participants appeared to be more accurate at timing accentuated onsets for impulsive sounds, but after correcting for repeated measures, this was not shown to be significant. Future research with a more dedicated experiment on onset timing could shed more light on this.

In summary, the hypothesis testing approach provides measures of the statistical significance of the differences between motion features for various groups. The method requires some initial hypothesis in order to determine which features to test and which groups to compare. Hypotheses can be formulated from knowledge gained e.g. by qualitative analysis of feature plots from the recordings.

6 Analysis III: Correlation Analysis

The hypothesis tests above indicated statistically significant differences between some features for some of the sound-tracings. However, the tests are not applicable to analyse the sound-motion links at a lower timescale level. To analyse correspondence between time-varying features of sound and motion, we apply a correlation analysis. As explained in Section 2.2, and pointed out by Schubert (2002), it is important to take into account the principle of *autocorrelation* when applying correlation to musical time series. The value of a sound feature at a certain point in time is likely to be close to the value of the same feature in the time period just before, and also the time period following it. For instance, a loud sound usually takes some time to decay. In the same way, we may assume that a participant’s bodily response to a sound stimulus at a certain time will not only be based on the sound parameters at that particular time, but it will also be based on memory of how the sound has unfolded so far, and expectation of how the sound will continue to change.

We have followed Schubert’s suggestion, and performed a Spearman’s ρ correlation analysis on the data series in our data set. For each recording, each of the 4 time-varying motion features was correlated to each of the 3 sound features, resulting in 12 correlation coefficients.

Statistics for the 12 correlation coefficients for all of the 528 recordings, showed that most of the coefficients were distributed over the full range [-1 1], with median close to zero. Figure 3 shows

that the exception was the high positive correlation coefficients between vertical position and pitch. For these coefficients, the median value was 0.87, while the medians of the other 11 coefficients were distributed between -0.02 and 0.38.⁴

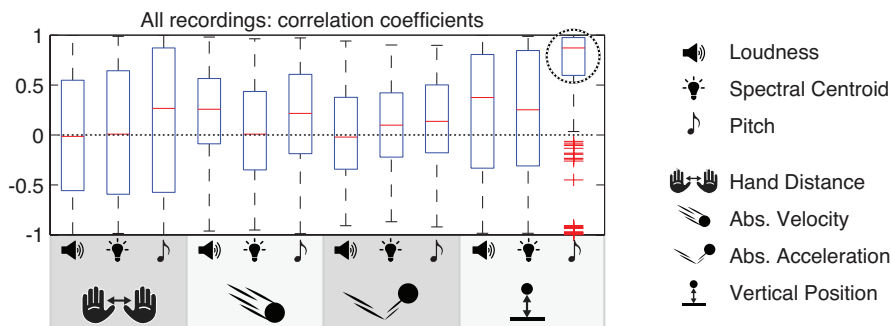


Figure 3: Correlation coefficients for all of the sound-tracings. The circled boxplot shows the correlation coefficients between pitch and vertical position. The red + marks are considered to be outliers.

By analysing the results from a single sound, we observed more clear tendencies in the correlations between sound and motion features. We display the analysis of sound-tracings related to Sound 1 as an example of detailed analysis. The other considerations provided in this section emerge from similar analysis of plots related to the other sounds. Plots related to Sound 1 are shown in Figure 4. The left column of this figure shows boxplots of the correlation coefficients, and the right column shows the same for the absolute values of the correlation coefficients, meaning that negative correlations are not distinguished from positive correlations. Three particularly interesting findings are marked in this figure. Point a) shows that overall, the correlation between pitch and vertical position was the strongest. In the sound features of Sound 1, there was a negative correlation between pitch and spectral centroid, which implies a negative correlation between vertical position and spectral centroid. The absolute values of the correlation coefficients for pitch were higher than for spectral centroid, indicating that pitch was the most salient of these sound features. Point b) shows that even though the correlation coefficients for the hand distance feature seem to be spread across the entire scale, the absolute values indicate that the correlations between distance and pitch were close to the extreme values (-1 and 1) for 9 out of 10 non-experts. This indicates that most non-experts followed the pitch feature with the distance between hands, but that they did not agree on whether to move their hands toward each other or away from each other. Note also that the absolute values of the correlation coefficients between hand distance and pitch are more in unison than the coefficients for vertical position and pitch. Point c) shows an example of what we observed in many of the sound-tracings: that the correlation coefficients from non-experts were more spread out than those from experts.

We compared features from the sound-tracings based on opposite sound feature envelopes against each other. For instance, Sound 1 (rising pitch and falling spectral centroid) was compared to Sound 2 (falling pitch and rising spectral centroid). Again, we observed a positive correlation between vertical position and pitch. For six out of the seven sounds with rising or falling pitch, there was a strong positive correlation. For Sound 15, which was both pitched and impulsive, the absolute values of these correlation coefficients were high. This indicates that the vertical position did follow the pitch envelope, but that the direction varied. An inspection of the sound-tracings of all the impulsive sounds (15, 16, 17 and 18), revealed that participants tended to put their hands down after making an initial impulsive gesture. Hence, the vertical position was correlated to the decreasing loudness envelope. The ambivalence in the correlation between vertical position and pitch of Sound 15 might be due to that some participants selected the impulsive quality of the sound as their main focus, and others picked up the rising pitch, and followed this with an upward hand motion.

As was stated for Sound 1 previously, we observed high correlation absolute values for the

⁴Readers may question why no test results are provided for the significance of the correlation coefficients. In this case, we remind the reader of Schubert's (2002) warning against applying significance tests on correlation coefficients from serial data.

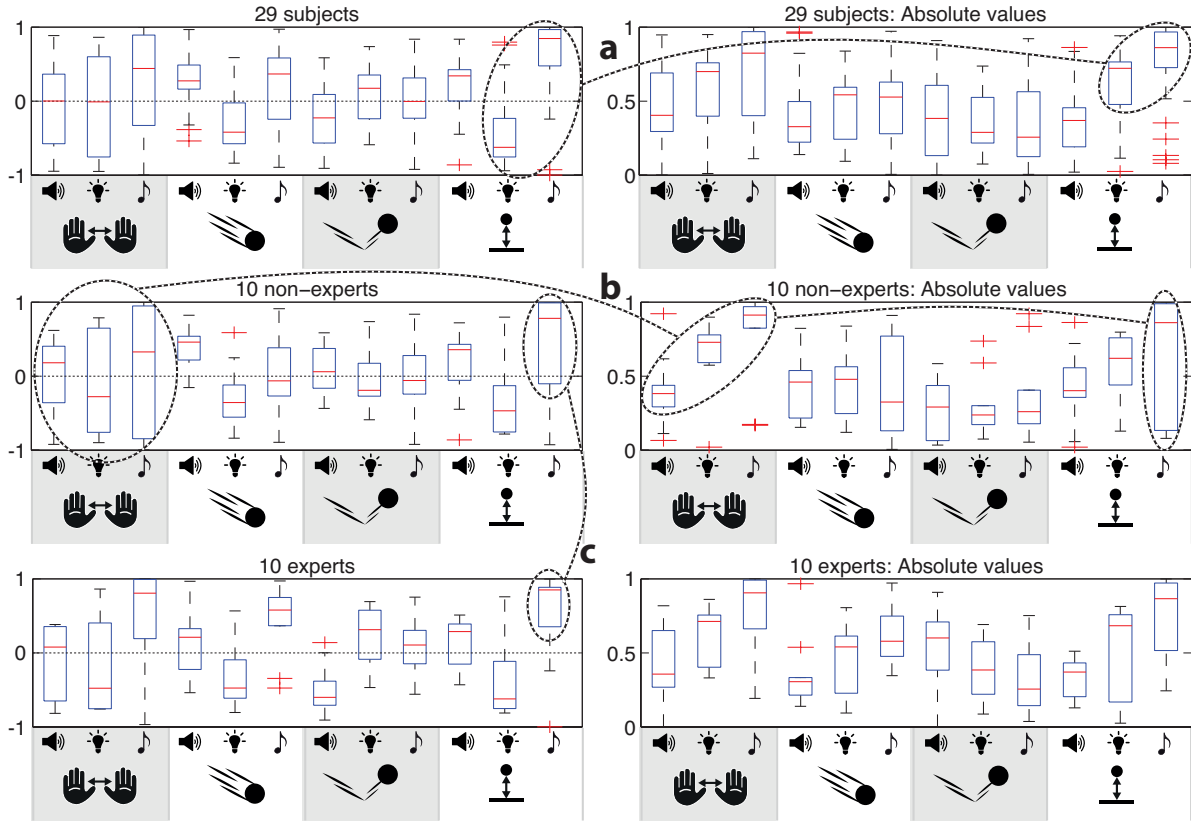


Figure 4: Correlation coefficients for the sound-tracings to Sound 1. Letters a, b and c show interesting results in the plot. The non-expert and expert groups display the results from the subjects that rated their level of musical training as *none/little* and *extensive*, respectively. Refer to the main text for details, and to Figure 3 for legend.

hand distance feature for several sound-tracings. However, the values were seemingly just as often positively correlated as negatively correlated to the various sound features. This suggests that participants often increased or decreased hand distance, but without considering a strategy for tracing a particular sound feature. We could speculate that this change in hand distance mimicked inputting energy to the sound, much like an accordion where the instrument makes sound regardless of the bellow being contracted or expanded.

For the sounds with a changing spectral centroid and no changing pitch, we observed a difference between rising and falling sounds. The sounds with a rising spectral centroid had a stronger positive correlation between vertical position and spectral centroid than the sounds with a falling spectral centroid. At the same time, the sounds with falling spectral centroid had higher absolute values for the correlation between hand distance and spectral centroid.

In summary, the Spearman’s ρ method is able to test the one-to-one relationship between a sound feature and a motion feature. However, due to the serial nature of the data, the coefficients are not applicable to significance testing, and the coefficients should not be understood as ‘literal’ descriptions of the relationship between the features. Nevertheless, detailed examination of the analysis results for each sound does provide indications of that certain sound-motion correlations are more salient than others.

7 Analysis IV: Canonical Correlation Analysis

While the Spearman’s ρ , discussed in the previous section, measured the relationship between two variables, *canonical correlation analysis* (CCA) is a technique for measuring the relationship between two *sets* of variables (Hotelling, 1936). This means that while the Spearman ρ considered one sound feature and one motion feature at a time, CCA analysis can be applied to analyse multivariate

dependencies between sound and motion. CCA has previously been applied by Caramiaux et al. (2010) and Nymoen et al. (2011a) in experiments analysing correspondences between sound features and motion features.

For each sound-tracing, the calculated time-varying features from the motion data (here called \mathbf{X}), and the sound features (\mathbf{Y}) were input to the CCA. In short, CCA then finds two vectors \mathbf{a} and \mathbf{b} which describe the linear combination of the motion features ($\mathbf{a}'\mathbf{X}$) and the linear combination of sound features ($\mathbf{b}'\mathbf{Y}$) that have the highest correlation with each other. The elements of \mathbf{a} and \mathbf{b} are called *canonical weights*, and $\mathbf{a}'\mathbf{X}$ and $\mathbf{b}'\mathbf{Y}$ are called the *first canonical components*. The next step is estimating another set of canonical weights \mathbf{a} and \mathbf{b} , optimizing the correlation between the *second canonical components* $\mathbf{a}'\mathbf{X}$ and $\mathbf{b}'\mathbf{Y}$, given the constraint that they are uncorrelated with the first canonical components. This process is repeated until the smallest number of features in \mathbf{X} or \mathbf{Y} is reached. Figure 5 illustrates this process for one sound file with one corresponding motion recording, each represented with only two features for the sake of clarity.

Several ways of interpreting the results of a canonical correlation analysis exist. We follow the approach of Caramiaux et al. (2010), and analyse the relationships between the sound and motion features by inspecting the *canonical loadings*, denoting the correlation between a feature and a corresponding canonical component.

Figure 6 shows statistics for the canonical loadings of all the sound-tracings. As outlined in this figure, the motion loadings for the first canonical component (which is the strongest component) show generally higher values for hand distance and vertical position than for absolute velocity and absolute acceleration. However, due to the issues regarding hypothesis testing of correlation coefficients from serial data, we can not assess the significance of this. Furthermore, the boxplots in the figure overlap partially, and what is more, not all sounds had a perceivable pitch. Hence, it is not sufficient to evaluate the overall results for all sounds in our experiment. Accordingly, we will look at how the individual canonical loadings are distributed.

A subset consisting of only the strongest canonical components was selected, only considering the sound-tracings of pitched sounds. A threshold was set at the lower quartile of the correlation coefficient for the first canonical components (0.94), and the components with correlation coefficients above this threshold were included in the analysis (307 components in all). We performed a simple classification of the canonical components in the subset, one based on the sound loadings and another on the motion loadings. Decision boundaries for the correlation coefficients were set manually at 0.5, and a canonical component was classified as correlated or not correlated to each feature.

The results of the analysis of individual canonical components indicated that distance between hands and vertical position were the most pertinent motion features in our data set. In most instances, two or three sound features were correlated to the corresponding canonical component, possibly because the sound features themselves were correlated to each other. We saw that for the canonical components that were only correlated to one sound feature, the sound feature was most often loudness. However, the reason might be that loudness was less often than pitch and spectral centroid correlated to other sound features. Furthermore, the pitch and spectral centroid envelopes were stable for some of the sounds. Hence, the only variance in these features would be due to noise, which again would lead to low correlation values.

In summary, CCA is useful for looking at how a set of sound features is related to a set of motion features. However, the analysis results are more difficult to interpret than the results from the previously presented methods. The multidimensionality of the CCA results made it necessary to simplify the presentation of the results through a decision boundary classification. Arguably, this has reduced some nuances in the results.

8 Discussion

After presenting the four different approaches for analysing the recorded motion data, we will now evaluate our results and the analysis methods used. We will start by looking at the results themselves in Section 8.1, and then continue by evaluating the utility of the analysis techniques in Section 8.2.

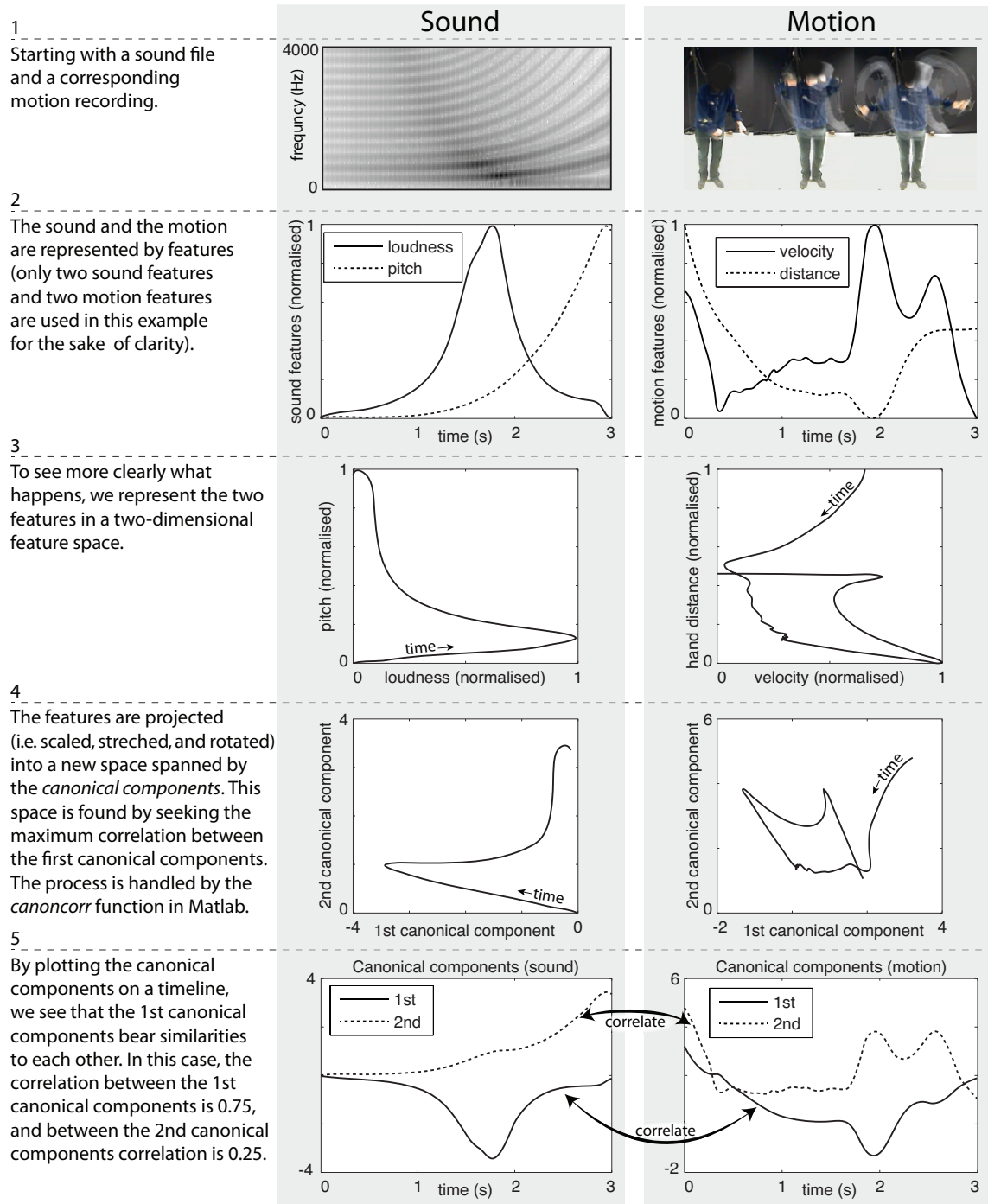


Figure 5: Simplified illustration of how CCA was applied to a single recording. Only two sound features and two motion features are used in this example. The features are projected onto the canonical weights, in order to obtain maximum correlation between the first canonical components. Subsequently (not shown in the figure), the correlation between each sound/motion feature and its corresponding canonical component is calculated, in order to determine the pertinence of each feature.

8.1 Results

The overall results from the pattern recognition analysis serve as a good point of departure for our discussion. This analysis confirmed that there were certain similarities between the sound-tracings that different subjects performed to the same sound. At a first glance, the precision of the classifier might not have been very impressive, with an average classification accuracy of 22.5%. However, if we would compare this accuracy to a random selection (which would give an accuracy of 5.6%),

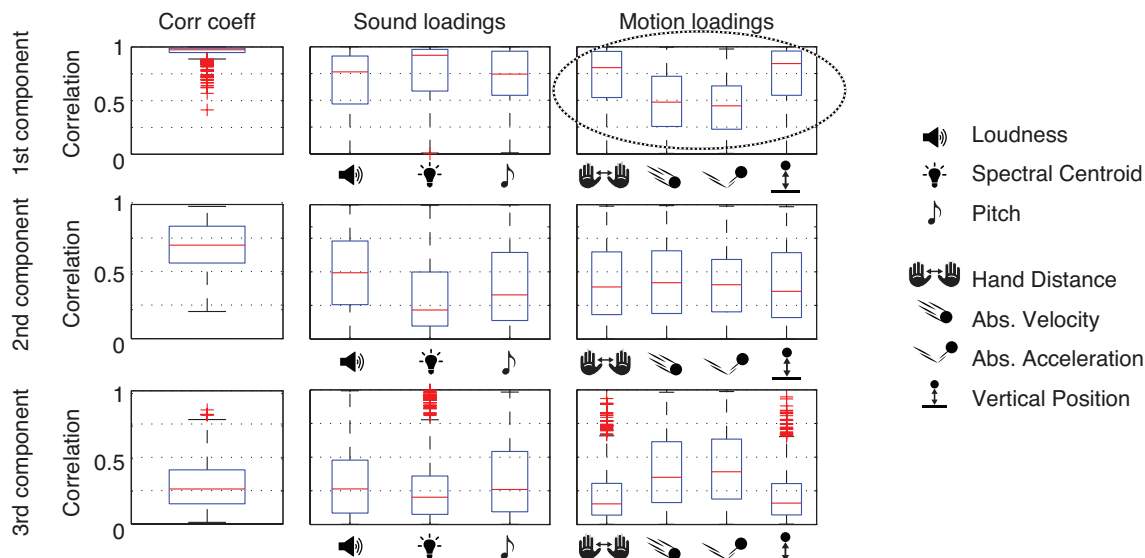


Figure 6: Box plots of the correlation strength and canonical loadings for all the sound-tracings in the experiment. Hand distance and vertical position have a noticeably higher pertinence on the first canonical component than the other motion parameters. This suggests that these two motion features are important when people try to relate motion to sound.

it is evident that the classifier indeed must have identified some similarities between the motion recordings. Furthermore, we could clearly see certain clusters in the output of the classifier. Most of the sound-tracings were predicted by the classifier to be one of the sound classes that were similar to the correct sound. This suggests that the extracted motion features reflected salient aspects of how we relate motion to sound.

The most apparent correspondence we observed between a sound feature and motion feature was the positive correlation between vertical position and pitch. This was not surprising, as the same has been reported in previous sound-tracing experiments by ourselves and other researchers, e.g. (Eitan and Granot, 2006; Huron et al., 2009; Nymo en et al., 2011a). This relationship may be explained through learned metaphors where pitch is represented in a vertical dimension, such as the vertical ordering of notes in a musical score, or the typical cartoon scenario where the bad coyote falls of a cliff, accompanied by a descending glissando.

It should be noted that another typical metaphor for pitch is the size metaphor, where high pitch would be related to small size (Huron et al., 2009). Hence, we could have expected a negative correlation between hand distance and pitch in our experiment. A high number of components in the CCA were correlated to both hand distance and pitch, and we also saw this relationship through high absolute values in the Spearman correlation. However, there was no agreement among the participants as to whether rising pitch should entail moving hands apart or together, and as such we cannot say that this is a strong relationship. Also, a *t*-test showed no significant difference between HandDivergenceMean for “falling” sounds and “rising” sounds.

The SVM classifier predicted several of the sounds with stable pitch but falling or rising spectral centroid (Sounds 7 and 8) to be one of the “steady” sounds, indicating that in addition to the sound-tracings of “falling” and “rising” sounds, where vertical position is a pertinent feature, these sound-tracings also had similarities with the sounds in the “steady” class. The same indication was previously found through an analysis of variance, where Sounds 7 and 8 caused significantly less vertical motion (Nymo en et al., 2012).

The classifier was able to distinguish quite well between sound-tracings of impulsive and non-impulsive sounds. Most of the sound-tracings to impulsive sounds had a high acceleration peak approximately synchronised with the sound onset. The significant difference in OnsetAcceleration between impulsive and non-impulsive sounds suggests that this feature was the most pertinent to distinguish between these groups.

8.2 Method

Having applied four different analysis techniques to the data set, we have observed conceptual differences between the methods, and also learned more about the strengths and weaknesses of each of them.

8.2.1 The Timescale Perspective

We start our discussion of the techniques by returning to the concept of *timescales* from Section 2.1.1. The Spearman ρ and CCA approaches both compare time-varying features, meaning that the analysis occurs at the *micro* timescale level. These methods can distinguish between different contours in the features. The *t*-tests and classifier methods cannot distinguish between such aspects unless global features reflecting all possible time-varying contours are extracted.

For correlation approaches, the participants' temporal accuracy in reproducing the contour of the sound feature may influence the correlation coefficient to a large extent. Such sensitivity to temporal accuracy may be desired if the goal is to evaluate how accurately a sound feature is followed, as done by Kussner (2012); but, if the goal is to identify relationships between sound and motion features, rapidly changing features must be reproduced with high temporal accuracy in order to be recognised at all. Admittedly, the features absolute velocity and absolute acceleration which were used in our analysis do change rapidly, and thus they might be more closely related to the sound features than what the CCA and Spearman ρ were able to identify. In principle, dynamic time warping (DTW) could be applied to compensate for this. However, since the timing of our study was quite controlled by using short sound objects and a countdown, we suspect that DTW would introduce more problems than it would solve. DTW would require preconceived assumptions of links between certain features, which again might undermine other, unknown sound-motion links.

The *t*-test and classifier methods are not necessarily as sensitive to rapid changes in the features. These methods both operate at the *meso* timescale. At this timescale, features can be designed to take possible temporal inaccuracies into account, such as the OnsetAcceleration feature, which was calculated as the mean acceleration in the time interval between 0.2 s before the sound onset and 0.5 s after the sound onset. Consequently, the feature captured any high acceleration peak in this time interval, regardless of whether or not the peak was perfectly synchronised with the sound onset.

If care is not taken to cover all of the relevant aspects in the feature extraction process, analysis at the meso timescale may fail to separate between different contours of the time-varying features. In other words, limiting the analysis to simple global features such as average values of time-varying features, could cause two completely different trajectories to appear equal. For instance, the mean vertical velocity of a fast downward motion followed by resting the hand in a fixed vertical position would be the same as if the motion was a slow downward trajectory ending up in the same vertical position. This could be compensated for by including features that take temporal aspects into account, for instance by looking at shorter time windows within the sound-tracing, or features based on non-linear transformations of the original time-varying features. In the present paper, OnsetAcceleration and VerticalEnergyUM are examples of such features.

8.2.2 Top-Down and Bottom Up

The analysis methods can also be evaluated in terms of being presumptive or naïve. The *t*-test method is a top-down approach in the sense that the researcher must make some assumption about the data set, and subsequently perform a quantitative test to see whether this assumption matches with the measurements. In contrast, the classifier, the Spearman correlation, and the CCA methods, are naïve approaches that do not require the same assumptions of the data set before the analysis is performed. These methods are bottom-up approaches, closer to data mining than to hypothesis testing. Thus, one might consider these methods as more “objective” than the *t*-test method; however, the process of extracting and selecting features is influenced by the researcher's assumptions about the data, and as such we cannot say that one is more objective than the other.

8.2.3 Level of Complexity

Another aspect of the analysis methods is their ability to analyse composite relationships between sound and motion. The Spearman ρ and the t -tests consider one motion feature and one sound feature at a time. The CCA and SVM methods should be applied to more than a single feature at a time, and are meant to find clusters or multivariate correlations in a data set. The SVM does not consider the features of the sounds *per se*, but rather the individual sound objects as classes. As was shown in this paper, the method can find clusters within a fairly large number of actions related to different sound objects. While the CCA results can be difficult to interpret if a large number of features are used, Caramiaux et al. (2010) have shown it to be useful for finding multivariate correlations for a smaller number of features.

8.2.4 Validity of the Results

Results from statistical tests provide strong indications of the validity of hypotheses put forward by researchers, and should be included whenever applicable. However, in experiments with continuous response data, this is often not the case. As argued by Schubert (2002), coefficients obtained from correlation of time-series should not be subjected to statistical tests. This does not entail that the assessments obtained from comparing correlation coefficients are invalid. Schubert's tests suggested that internal validity could be claimed for Pearson correlation coefficients above 0.75. Schubert does not provide an equivalent threshold for Spearman ρ , however, since Spearman ρ should be a better measure for correlation of musical time-series, it is reasonable to assume that the corresponding threshold is lower (Schubert, 2002).

9 Conclusions and Future Work

We have shown that different techniques for analysing sound-tracings have different strengths and weaknesses. The choice of method depends on the aim of the research, but when sound-motion links at both meso and micro timescales are of interest, it may be appropriate to apply multiple analysis methods. Our analysis showed that people performed similar actions to sounds with rising tendencies, and also to sounds with falling tendencies, whether the rise or fall was in pitch or in spectral centroid. These similarities existed even when the sounds were easy to tell apart. As such, we can say that the sounds in our experiment were more detailed than the motion responses.

Although this paper has shed light upon some of the possible strategies for quantitative analysis of sound-tracings, there is still a need for defining methods that can analyse correspondences between input data and response data that are time-varying and multidimensional, and that ideally also allows proper testing of the significance of the results. We find the challenge of analysing multidimensional data in music and motion intriguing and hope that the challenge will inspire readers to contribute to future research in this area.

The experiment in the present paper did not consider the spatial aspect of sound. Our lab has recently acquired a 32-channel speaker setup, and we plan to pursue research on links between sound spatialisation and body motion in the future. In future work we will also apply the results of our sound-tracing studies in choosing algorithms for actively controlling music in mobile music systems, and to define mappings in new musical interfaces.

Furthermore, we plan to continue our experiments on cross-modal analysis, specifically by having a different set of subjects match the sounds that were used in this experiment to the recorded motion data.

10 Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement no. 257906, Engineering Proprioception in Computer Sys-

tems (EPiCS), and the Norwegian Research Council, project no. 183180 Sensing Music-Related Actions (SMA).

The authors would like to thank Pavel Zahorik, Daniel Leech-Wilkinson, and two anonymous reviewers for insightful comments in the preparation of this manuscript.

References

- BREGMAN, A. S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA.
- BURGER, B., THOMPSON, M., LUCK, G., SAARIKALLIO, S., AND TOIVIAINEN, P. 2012. Music moves us: Beat-related musical features influence regularity of music-induced movement. In *Proceedings of the International Conference on Music Perception and Cognition*. Thessaloniki, 183–187.
- CAMURRI, A., LAGERLÖF, I., AND VOLPE, G. 2003. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies* 59, 1–2, 213–225.
- CARAMIAUX, B., BEVILACQUA, F., AND SCHNELL, N. 2010. Towards a gesture-sound cross-modal analysis. In *Gesture in Embodied Communication and Human-Computer Interaction*, S. Kopp and I. Wachsmuth, Eds. Lecture Notes in Computer Science Series, vol. 5934. Springer, Berlin / Heidelberg, 158–170.
- CLARKE, E. F. 2005. *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford University Press, New York.
- DUDA, R. O., HART, P. E., AND STORK, D. G. 2000. *Pattern Classification*. Wiley-Interscience, New York.
- EITAN, Z. AND GRANOT, R. Y. 2006. How music moves: Musical parameters and listeners’ images of motion. *Music Perception* 23, 3, 221–248.
- GALANTUCCI, B., FOWLER, C., AND TURVEY, M. 2006. The motor theory of speech perception reviewed. *Psychonomic bulletin & review* 13, 3, 361–377.
- GODØY, R. I. 2006. Gestural-sonorous objects: Embodied extensions of Schaeffer’s conceptual apparatus. *Organised Sound* 11, 02, 149–157.
- GODØY, R. I. 2010a. Gestural affordances of musical sound. In *Musical Gestures: Sound, Movement, and Meaning*, R. I. Godøy and M. Leman, Eds. Routledge, New York, Chapter 5, 103–125.
- GODØY, R. I. 2010b. Images of sonic objects. *Organised Sound* 12, 54–62.
- GODØY, R. I., HAGA, E., AND JENSENIUS, A. R. 2006. Exploring music-related gestures by sound-tracing, a preliminary study. In *2nd ConGAS International Symposium on Gesture Interfaces for Multimedia Systems*. Leeds, UK.
- GODØY, R. I. AND JENSENIUS, A. R. 2009. Body movement in music information retrieval. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*. Kobe, Japan, 45–50.
- GREY, J. M. 1977. Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.* 61, 5, 1270–1277.
- HAGA, E. 2008. Correspondences between music and body movement. Ph.D. thesis, University of Oslo.
- HOLM, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.
- HOTELLING, H. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4, 321–377.
- HURON, D., DAHL, S., AND JOHNSON, R. 2009. Facial expression and vocal pitch height: Evidence of an intermodal association. *Empirical Musicology Review* 4, 3, 93–100.
- KOHLER, E., KEYSERS, C., UMITÀ, A., FOGASSI, L., GALLESE, V., AND RIZZOLATTI, G. 2002. Hearing sounds, understanding actions: Action representation in mirror neurons. *Science* 297, 5582, 846–848.

- KOZAK, M., NYMOEN, K., AND GODØY, R. I. 2011. The effects of spectral features of sound on gesture type and timing. In *Proc. of the 9th International Gesture Workshop*, E. Efthimiou and G. Kouroupetroglou, Eds. Athens, 20–24.
- KUSSNER, M. 2012. Creating shapes: musicians’ and non-musicians’ visual representations of sound. In *Proceedings of 4th Int. Conf. of Students of Systematic Musicology*, U. Seifert and J. Wewers, Eds. epOs-Music, Osnabrück (Forthcoming).
- LARTILLOT, O., TOIVAINEN, P., AND EEROLA, T. 2008. A matlab toolbox for music information retrieval. In *Data Analysis, Machine Learning and Applications*, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin / Heidelberg, 261–268.
- LEMAN, M. 2008. *Embodied Music Cognition and Mediation Technology*. The MIT Press, Cambridge, MA.
- LEMAN, M., VERMEULEN, V., DE VOOGDT, L., TAELEMAN, J., MOELANTS, D., AND LESAFFRE, M. 2004. Correlation of gestural musical audio cues and perceived expressive qualities. In *Gesture-Based Communication in Human-Computer Interaction*, A. Camurri and G. Volpe, Eds. Lecture Notes in Computer Science Series, vol. 2915. Springer, Berlin / Heidelberg, 40–54.
- LIBERMAN, A. M. AND MATTINGLY, I. G. 1985. The motor theory of speech perception revised. *Cognition* 21, 1, 1 – 36.
- MCADAMS, S., WINSBERG, S., DONNADIEU, S., SOETE, G., AND KRIMPHOFF, J. 1995. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research* 58, 177–192.
- MERER, A., YSTAD, S., KRONLAND-MARTINET, R., AND ARAMAKI, M. 2008. Semiotics of sounds evoking motions: Categorization and acoustic features. In *Computer Music Modeling and Retrieval. Sense of Sounds*, R. Kronland-Martinet, S. Ystad, and K. Jensen, Eds. Lecture Notes in Computer Science Series, vol. 4969. Springer-Verlag, Berlin, Heidelberg, 139–158.
- MIERSWA, I., WURST, M., KLINKENBERG, R., SCHOLZ, M., AND EULER, T. 2006. Yale: Rapid prototyping for complex data mining tasks. In *KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, Eds. ACM, New York, 935–940.
- NYMOEN, K., CARAMIAUX, B., KOZAK, M., AND TORRESEN, J. 2011a. Analyzing sound tracings: a multimodal approach to music information retrieval. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*. MIRUM ’11. ACM, New York, 39–44.
- NYMOEN, K., GLETTE, K., SKOGSTAD, S. A., TORRESEN, J., AND JENSENIUS, A. R. 2010. Searching for cross-individual relationships between sound and movement features using an SVM classifier. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Sydney, 259–262.
- NYMOEN, K., SKOGSTAD, S. A., AND JENSENIUS, A. R. 2011b. Soundsaber - a motion capture instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Oslo, 312–315.
- NYMOEN, K., TORRESEN, J., GODØY, R., AND JENSENIUS, A. R. 2012. A statistical approach to analyzing sound tracings. In *Speech, Sound and Music Processing: Embracing Research in India*, S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen, and S. Mohanty, Eds. Lecture Notes in Computer Science Series, vol. 7172. Springer, Berlin Heidelberg, 120–145.
- PEETERS, G., GIORDANO, B., SUSINI, P., MISDARIIS, N., AND MCADAMS, S. 2011. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America* 130, 5, 2902–2916.
- PELLEGRINO, G., FADIGA, L., FOGASSI, L., GALLESE, V., AND RIZZOLATTI, G. 1992. Understanding motor events: a neurophysiological study. *Experimental Brain Research* 91, 176–180.
- PIRHONEN, A. 2007. Semantics of sounds and images – can they be paralleled? In *Proceedings of the 13th International Conference on Auditory Display*. Montreal, 319–325.

- REPP, B. H. 1995. Quantitative effects of global tempo on expressive timing in music performance: Some perceptual evidence. *Music Perception* 13, 1, 39–57.
- ROSENBAUM, D. 2001. *Human motor control*. Academic Press, San Diego.
- SCHAEFFER, P. 1966. *Traité des objets musicaux*. Éditions du Seuil, Paris.
- SCHUBERT, E. 2002. Correlation analysis of continuous emotional response to music. *Musicae Scientiæ Spec. Is.* 2001-02, 213–236.
- STEIN, B. E. AND MEREDITH, M. A. 1993. *The merging of the senses*. The MIT Press, Cambridge, MA.
- STYNS, F., VAN NOORDEN, L., MOELANTS, D., AND LEMAN, M. 2007. Walking on music. *Hum Movement Sci* 26, 5, 769–785.
- VAN NOORDEN, L. 1975. Temporal coherence in the perception of tone sequences. Ph.D. thesis, Technical University Eindhoven.
- VAN NORT, D. 2009. Instrumental Listening: Sonic gesture as design principle. *Organised Sound* 14, 02, 177–187.
- VASSILAKIS, P. 2001. Perceptual and physical properties of amplitude fluctuation and their musical significance. Ph.D. thesis, University of California, Los Angeles.
- VINES, B. W., KRUMHANSL, C. L., WANDERLEY, M. M., AND LEVITIN, D. J. 2006. Cross-modal interactions in the perception of musical performance. *Cognition* 101, 1, 80–113.
- VROOMEN, J. AND DE GEDLER, B. 2000. Sound enhances visual perception: Cross-modal effects on auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance* 26, 5, 1583–1590.
- WALKER, B. N. 2000. Magnitude estimation of conceptual data dimensions for use in sonification. Ph.D. thesis, Rice University, Houston, TX.
- WALKER, R. 1987. The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception & Psychophysics*, 491–502.
- WESSEL, D. L. 1979. Timbre space as a musical control structure. *Computer Music Journal* 3, 2, 45–52.
- ZATORRE, R. 2005. Music, the food of neuroscience? *Nature* 434, 312–315.