Searching for Cross-Individual Relationships between Sound and Movement Features using an SVM Classifier

Kristian Nymoen, Kyrre Glette, Ståle Skogstad, Jim Torresen, Alexander R. Jensenius[†]

University of Oslo
Department of Informatics,
Pb 1080 Blindern, 0316 Oslo, Norway
{krisny, kyrrehg, savskogs, jimtoer}@ifi.uio.no

[†]University of Oslo
Department of Musicology
Pb 1017, Blindern, 0315 Oslo, Norway
a.r.jensenius@imv.uio.no

ABSTRACT

In this paper we present a method for studying relationships between features of sound and features of movement. The method has been tested by carrying out an experiment with people moving an object in space along with short sounds. 3D position data of the object was recorded and several features were calculated from each of the recordings. These features were provided as input to a classifier which was able to classify the recorded actions satisfactorily, particularly when taking into account that the only link between the actions performed by the different subjects was the sound they heard while making the action.

1. INTRODUCTION

What are the underlying links between movement and sound? We believe that the way we perceive sounds and their sound-producing actions are related, and that this relationship may be explored by observing human movement to sound. Auditory sensations are often perceived as mental images of what caused the sound. This idea of a *gestural-sonic object* is built upon motor theory in linguistics and neuroscience [8]. This belief has motivated an experiment to explore how sound and body movement are related: Is it possible to discover cross-individual relationships between how we perceive features of sound and features of movement by studying how people choose to move to sounds? The term *cross-individual* here denotes relationships that are found in the majority of the subjects in this experiment. Further, we use *movement* to denote continuous motion, and *action* to denote a segment of motion data.

Several papers have focused on training a machine learning system to recognize a specific action. This paper, however, presents a technique for discovering correlations between sound features and movement features. We investigate the use of a machine learning system to classify the actions that subjects link to certain sounds, here denoted as *sound-tracings* [9]. The features used for classification are evaluated, and the results of presenting various subsets of those features to the classifier are explored. This makes it possible to discover how a classification of sound-tracings based on certain *action* features is able to distinguish between *sounds* with certain characteristics. At the same time the classifier may be unable to distinguish between sounds with other characteristics. For instance, one of our hypotheses has been that features related to velocity would distinguish well between sounds with different loudness envelopes. Another hypothesis is that the features related to vertical displacement would distinguish between sounds with different pitch envelopes. An analysis of the classifier's performance can provide information on natural relationships between sounds and actions. This is valuable information in our research on new musical instruments.

Section 2 gives a brief overview of related research, including some notes on previous use of machine learning to classify music-related movement. Section 3 gives an overview of the method used. Section 4 presents the classification of the data, including feature extraction from the movement data and some results on reducing the number of inputs to the classifier. Finally, in section 5 we discuss the method used in the light of the results presented in section 4, and provide some conclusions and plans for future work on this material.

2. RELATED WORK

Machine learning and pattern recognition of motion data have been applied in musical contexts in various ways. Early works on applying neural networks to recognize actions to be mapped to sound synthesis parameters were presented in the early 1990s [5, 10]. In the last decade, various other machine learning implementations of mapping motion capture data to sound synthesis have been presented. This includes toolkits for machine learning in PureData [3] and Max/MSP [1], and a tool for on-the-fly learning where the system is able to learn new mappings, for instance during a musical performance [6].

Although mapping applications seem to have been the most used implementation of machine learning on motion data in musical contexts, some analytical applications exist as well. In *EyesWeb*, Camurri et al. have implemented recognition of expressivity in what they call 'musical gestures' [2]. Machine learning has also been applied to instrumental actions, like extraction of bowing features and classification of different bow strokes in violin performance [12, 13].

A significant amount of work has been done on information retrieval of motion capture data within research fields related to computer animation [11]. Much of the work in this field has been on classification of different actions in a motion database (e.g. distinguishing a kicking action from a jumping action). For this sort of classification Müller and Röder have introduced *motion templates* [11]. This method is based on spatio-temporal relationships between various parts of the body. They present a sophisticated method for recognizing specific actions, a method which is independent

NIME2010, June 15-18, 2010, Sydney, Australia

from numerical differences in the raw data.

The research presented in this paper distinguishes itself from the previously mentioned ones in that it aims to recognize certain unknown features of the actions rather than the actions themselves. The approach is analytical, with a goal of discovering cross-individual relationships between features of sound and features of movement.

A similar experiment to the one presented in this paper was carried out in 2006, where subjects were presented with short sounds and instructed to sketch sound-tracings on a Wacom tablet [9]. This data was initially studied qualitatively, and has recently also been processed quantitatively in an unpublished manuscript which inspired this paper [7].

3. METHOD

3.1 Setup

In our experiment we used a 7 camera *Optitrack* infrared motion capture system for gathering position data of reflective markers on a rod. A sampling rate of 100 Hz was used, and data was sent in real-time to Max/MSP for recording.

3.2 Observation Experiment

Fifteen subjects, with musical experience ranging from no performance experience to professional musicians, were recruited. These were 4 females and 11 males, selected among university students and staff. The subjects were presented with ten sounds and asked to move a rod in space along with each sound, as if they themselves were creating the sound. The rod was roughly 120 cm long with a diameter of 4 cm (Figure 1). Before recording the movement data, the subjects listened to the sound twice (or more if they requested it), to allow them to make up their mind on what they thought would be a natural connection between the sound and the movement. A metronome was used so that the subjects could know at what time the sound started. The motion capture recording started 500 ms before the sound, and was stopped at the end of the sound file. Thus, all the motion capture recordings related to a single sound were of equal length which made it easier to compare the results from different subjects. We made three recordings of each action from each subject. Some of the recordings were discarded, due to the subject moving the rod out of the capture volume, which caused gaps in the data. Hence, there are between 42 and 45 data recordings of actions performed to each sound.

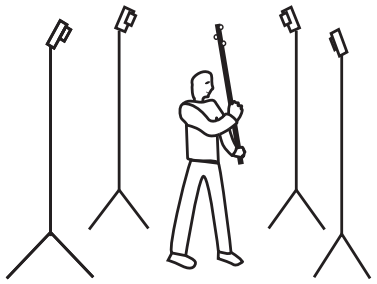

Figure 1: The figure shows a subject holding the rod with reflective markers in one end. Motion capture cameras are surrounding the subject.

The recorded data was the 3D position at the end of the rod, in addition to video. The subjects also filled out a small questionnaire where they were asked whether they considered themselves to be novices, intermediates or music experts, and whether they found anything in the experiment to be particularly difficult.

3.3 Sounds

The sounds used in the experiment all had one or more distinct features (e.g. rising pitch or varying sound intensity), which we believed would make the users move differently to the different sounds. A brief overview of the sounds is presented in Table 1, and the sounds are available online.[1] Some of the sounds were quite similar to each other, e.g. with only subtle differences in the timing of loudness peaks. As we shall see, actions performed to these similar sounds were often mistaken for each other by the classifier. Sounds 1 and 2 are similar, where the loudness and the center frequency of a bandpass filter sweeps up and down three times. The difference between the sounds is the timing of the peaks, which gives a slightly different listening experience. Sounds 9 and 10 are also quite similar to each other, with the same rhythmic pattern. The difference between the two is that Sound 9 has a steady envelope, while Sound 10 has impulsive attacks with a decaying loudness envelope after each attack.

Table 1: Simple description of the sounds used in the experiment

Sound	Pitch	Spectral Centroid	Loudness	Onsets
1	Noise	3 sweeps	3 sweeps	3
2	Noise	3 sweeps	3 sweeps	3
3	Falling	Rising	Steady	1
4	Rising	Falling	Steady	1
5	Noise	Rising	Steady	1
6	Noise	Rising / Complex	Steady	1
7	Noise	Rising, then falling	Steady	1
8	Rising	Complex	Steady	1
9	Noise	Steady	Rhythm:♩♩♫♩ Static (on/off)	5
10	Noise	Complex	Like 9, with decaying slopes	5

3.4 Software

For classification we used *RapidMiner*,[2] a user-friendly toolbox for data mining, classification and machine learning. A brief test of the various classification algorithms in RapidMiner indicated that Support Vector Machines (SVM) would provide the highest classification accuracies, so this was chosen for the experiments. RapidMiner uses the LIBSVM[3] library for SVMs. The python-script *grid.py* is provided with LIBSVM and was used for finding the best parameters for the algorithm. This script performs a simple grid search to determine the best parameters.

When training and validating the system, *cross-validation* was used due to the limited number of data examples. This means that two complementary subsets are randomly generated from the full data set. One subset of the data examples is used for training the classifier, and the other is used as a validation set to measure the performance of the classifier [4]. This process was repeated ten times with different subsets. Finally, the performance results ere averaged across all performance evaluations. Matlab was used for preprocessing and feature extraction.

4. ANALYSIS

The analysis process consists of two main parts: the feature extraction and the classification. In our opinion, the former is the most interesting in the context of this paper, where the goal is to evaluate a method for comparing movement features to sound features. The features selected are features that we believed would distinguish between the sounds.

[1] http://folk.uio.no/krisny/nime2010/
[2] http://rapid-i.com/
[3] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

4.1 Feature Extraction

When extracting the movement features, it is important to note that two actions that seem similar to the human eye do not need to be similar numerically. This implies that the features should be based on relative, rather than absolute data. In our setup, we have a recording of only a single point in space, and thus we cannot calculate spatial relations as suggested by Müller et al.[11], but we can look at temporal relations. Hence, we have chosen to base the features presented here on time-series based on the derivative of the position data. Since we have no reference data on the position of the subject (only the rod), we cannot tell whether horizontal movement is forwards or sideways. Thus horizontal movement along either of the two horizontal axes should be considered as equivalent. However, the vertical component of the movement can be distinguished from any horizontal movement, since gravity is a natural reference.

The 3D position data was used to calculate the following features from the recorded data:

- *VelocityMean* and *VelocityStD* are the mean and standard deviation of the vector length of the first derivatives of the 3D position data.
- *AccelerationMean* is the mean value of the vector length of the second derivative of the 3D position data.
- *TurnMean* and *TurnStD* are the mean value and the standard deviation of change in direction between the samples, i.e. the angle between the vector from sample n to n+1, and the vector from n+1 to n+2.
- *PreMove* is the cumulative distance before the sound starts. This is a period of 50 samples in all recordings.
- *vVelocityMean* is the mean value of the derivatives of the vertical axis. As opposed to *VelocityMean*, this feature can have both positive (upwards) and negative (downwards) values.
- *vEnergy* is an exponentially scaled version of *vVelocityMean*, meaning that fast movement counts more than slow movement. For example, fast movement downwards followed by slow movement upwards would generate a negative value, even if the total distance traveled upwards and downwards is the same.

Finally, each recording was divided into four equally sized segments, e.g. to be able to see how the first part of the action differed from the last part. The variables *segmentVelMean* — the mean velocity of each segment, and *segmentShake* — a measure based on autocorrelation to discover shaking, were calculated.

In the next section we will present the classification results, and investigate if classifications based on different subsets of features will reveal relationships between sound features and action features.

4.2 Results

When all the movement features were fed to the classifier, a classification accuracy of 78.6% ± 7.3% was obtained. This should be interpreted as the precision of recognizing the *sound* that inspired a certain action, based on features extracted from the *movement* data. Sound 7 was the one with the best accuracy, where the algorithm classified the 95.2% of the actions correctly, as shown in Table 2. The classifier misinterpreted some of the actions made to similar sounds, but still the lowest individual classification accuracy was as high as 68.9%. The table columns show the true actions, and the rows show the predictions of the classifier. The diagonal from top left to lower right indicates the correctly classified instances (marked in grey). We define *class recall (CR)* and *class precision (CP)* of class i as:

$$CR_i = \frac{||R_i \cap A_i||}{||R_i||} * 100\% \qquad CP_i = \frac{||R_i \cap A_i||}{||A_i||} * 100\%$$

$||A_i||$ denotes the number of instances classified as i, and $||R_i||$ denotes the total numbers of instances in class i. Then CP is the probability that a certain prediction made by the classifier is correct, and CR is the probability that the classifier will provide the correct result, given a certain class.

When reducing the features fed to the classifier to only include the two features related to vertical displacement, i.e. *vVelocityMean* and *vEnergy*, the total classification accuracy was reduced to 36%. However, the sounds with a distinct rising or falling pitch had significantly less change in classification accuracy than other sounds. For Sounds 3 and 4, we obtained a class recall of 79.1% and 51.2%, respectively. In addition to this we obtained a class recall of

Table 2: Classification accuracies for the individual sounds, when using all sound features. CP and CR denote class precision and class recall in percent, respectively. t1–t10 are the true classes, p1–p10 are the predictions made by the classifier.

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	CP
p1	34	6	1	1	1	0	0	0	0	4	72.3
p2	9	36	0	0	0	1	0	2	0	0	75.0
p3	0	0	36	2	0	2	0	0	0	0	90.0
p4	0	0	2	32	1	0	1	3	0	0	82.1
p5	0	0	1	2	31	6	1	2	1	0	70.5
p6	1	0	3	0	6	32	0	1	2	0	71.1
p7	0	0	0	0	1	0	40	3	0	0	90.9
p8	1	0	0	6	3	1	0	34	0	0	75.6
p9	0	1	0	0	2	2	0	0	36	6	76.6
p10	0	0	0	0	0	0	0	0	6	34	85.0
CR	75.6	83.7	83.7	74.4	68.9	72.7	95.2	75.6	80.0	77.3	

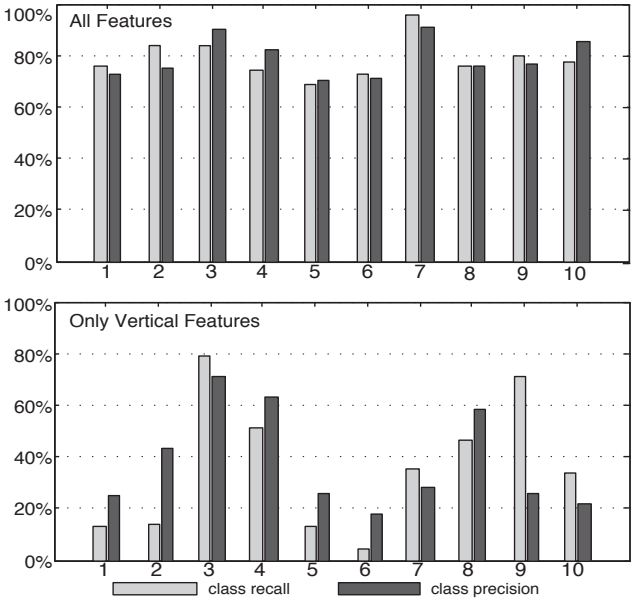

Figure 2: The figure shows the *class precision* and *class recall* for each of the classes (see text for explanation). A class consists of sound-tracings related to the same sound. High scores on both bars indicate that the estimation of this class is satisfactory. In the top chart, all features have been used as input to the classifier, in the lower chart, only the vertical displacement features were used.

71.1% for Sound 9, however the class precision of this sound was as low as 25.8%, indicating that the SVM classifier has made the class too broad for this to be regarded truly significant. The lower plot in Figure 2 shows that the class recall and class precision for all the sounds with changing pitch (3, 4 and 8) have relatively high scores on both accuracy and precision.

5. DISCUSSION

Our goal in this paper is to evaluate the use of a classifier to discover correlations between sound features and movement features. We have evaluated the method by using the data related to Sound 3, where we discovered a relationship between pitch and vertical movement. The fundamental frequency of this sound decreases from 300 Hz to 200 Hz. Figure 3 shows the vertical components of the actions performed to this sound by the subjects. The heavy lines denote the mean value and standard deviation of the vertical positions. Some of the actions do not follow the pitch downwards. This may be because the subject chose to follow the upwards moving spectral centroid. Also, quite a few of the actions make a small trip upwards before moving downwards. Still, there is a clear tendency of downwards movement in most of the performances, so we believe it is safe to conclude that there is a relationship between pitch and vertical position in our dataset. This finding makes it interesting to study the relationship between vertical position and pitch in a larger scale. Would we find similar results in a group that is large enough for statistical significance? Further on, we might ask if this action-sound relationship depends on things like cultural background or musical training.

We have also found similar, although not equally strong, indications of other correlations between sound and movement features. One such correlation is the *shake* feature. With only this as input, the classifier was able to distinguish well between Sounds 9 and 10. These were two rhythmic segments where the only difference was that Sound 10 had decaying slopes after each attack and Sound 9 had simply sound on or sound off with no adjustments in between. This could indicate that for one of the sounds, the subjects performed actions with impulsive attacks, resulting in a rebound effect which has been picked up in the shake feature.

Another relationship is the features *turnMean* and *turnStD* which seem to distinguish between the number of onsets in the sound. Sounds 1 and 2 had three onsets, and were quite well distinguished from the rest, but often confused with each other. The same was the case for Sounds 3, 4, 5, 6, 7 and 8 which had a single onset and Sounds 9 and 10 which had five onsets. A plausible explanation for this is that the subjects tended to repeat the same action for each onset of the sound, implying a somewhat circular movement for each onset. This circular motion is picked up in *TurnMean* and *TurnStD*.

The relationship between pitch and vertical displacement described in this section may seem obvious. But we believe the method is the most interesting. By using a classifier, we get an idea of where to look for cross-individual correlations between sound features and movement features.

6. CONCLUSIONS AND FUTURE WORK

The paper has presented a method for studying how perception of sound and movement is related. We believe that machine learning techniques may provide good indications of cross-individual correlations between sound features and movement features. Our experiments have shown that it is possible to study these relationships by feeding movement data to a classifier and carefully selecting the features used for classification. The paper has mainly focused on evaluating the method itself rather than the results, since a larger selection of subjects would be necessary to draw strong conclusions on the existence of action-sound relationships. Future research plans include experiments with a larger selection of subjects, and to expand the setup to full-body motion capture. In our research, we hope to learn more about how features of movement can be used to develop new intuitive movement-based instruments.

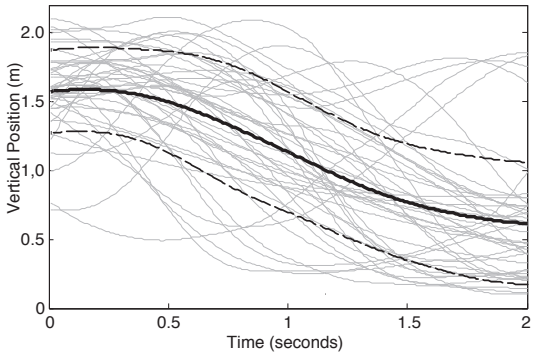

Figure 3: Plot of vertical position of the performances to Sound 3. The heavy lines denote mean value and standard deviation.

7. REFERENCES

[1] F. Bevilacqua, R. Müller, and N. Schnell. MnM: a Max/MSP mapping toolbox. In *Proceedings of NIME 2005*, pages 85–88, Vancouver, BC, 2005.

[2] A. Camurri, B. Mazzarino, and G. Volpe. Analysis of expressive gesture: The Eyesweb expressive gesture processing library. *Lecture Notes in Computer Science*, 2915:460–467, February 2004.

[3] A. Cont, T. Coduys, and C. Henry. Real-time gesture mapping in PD environment using neural networks. In *Proceedings of NIME 2004*, pages 39–42, Singapore, 2004.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[5] S. Fels and G. Hinton. Glove-talk: A neural network interface between a dataglove and a speech synthesizer. *IEEE Trans. Neural Networks*, 4(1):2–8, 1993.

[6] R. Fiebrink, D. Trueman, and P. R. Cook. A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of NIME 2009*, Pittsburgh, 2009.

[7] K. Glette. Extracting action-sound features from a sound-tracing study. Tech report, University of Oslo, 2009.

[8] R. I. Godøy. Gestural-sonorous objects: embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound*, 11(02):149–157, 2006.

[9] 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.

[10] M. Lee, A. Freed, and D. Wessel. Neural networks for simultaneous classification and parameter estimation in musical instrument control. *Adaptive and Learning Systems*, 1706:244–255, 1992.

[11] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/ SCA '06:*, pages 137–146. Eurographics Association, 2006.

[12] N. H. Rasamimanana, E. Fléty, and F. Bevilacqua. Gesture analysis of violin bow strokes. In S. Gibet, N. Courty, and J.-F. Kamp, editors, *Gesture Workshop*, volume 3881 of *LNCS*, pages 145–155. Springer, 2005.

[13] E. Schoonderwaldt and M. Demoucron. Extraction of bowing parameters from violin performance combining motion capture and sensors. *The Journal of the Acoustical Society of America*, 126(5), 2009.