Norwegian Artificial Intelligence Symposium, Gjøvik, 22 November 2010

Extracting Action-Sound Features From a Sound-Tracing Study

Kyrre Glette
University of Oslo, Department of Informatics
P.O. Box 1080 Blindern, 0316 Oslo, Norway
kyrrehg@ifi.uio.no

Alexander Refsum Jensenius, Rolf Inge Godøy
University of Oslo, Department of Musicology
P.O. Box 1017 Blindern, 0315 Oslo, Norway
{a.r.jensenius, r.i.godoy}@imv.uio.no

***Abstract*—The paper addresses possibilities of extracting information from music-related actions, in the particular case of what we call sound-tracings. These tracings are recordings from a graphics tablet of subjects' drawings associated with a set of short sounds. Although the subjects' associations to sounds are very subjective, and thus the resulting tracings are very different, an attempt is made at extracting some global features which can be used for comparison between tracings. These features are then analyzed and classified with an SVM classifier.**

1. Introduction

Navigation and search in music information retrieval (MIR) systems often focus on using either verbal descriptors or sonic features as input parameters. From our studies of peoples' spontaneous body movement to musical sound [3], [6], [5], [7], we believe there is a large, and largely undiscovered, potential in using body movement in navigation and search of sound and music collections. In our effort to develop such systems we are currently exploring machine learning techniques that can help in extracting relevant features from music-related body movement, musical sound and the relationships between movement and sound.

In this paper we report on a machine learning system for extracting relevant features from data of subjects' spontaneous "drawing" of short sounds. This we call "sound-tracing" and can be seen as a way of sketching salient features in the perceived sound. Just in the same way as sketching of visual material (e.g. cartoons) tend to emphasize some salient features of the people or objects being drawn, we believe that a similar type of approximate rendition of actions corresponding to some perceptually salient features in the musical sound may be seen in sound-tracing.

To test this in practice, we carried out a pilot experiment where subjects were asked to sketch quickly some features they associated with the sound. A qualitative analysis of the material was presented in [5], and here we will present a machine-learning approach to the same material. The main interest in the sound-tracing study was to see what types of sonic features the subjects would respond to, and how they would trace these features with the digital pen, for instance:

- Mimicking the sound-producing action(s), e.g. impulsive (short burst of effort followed by relaxation), sustained (continuous effort) and iterative (a rapid series of small and/or back-and-forth movements).
- Tracing features in the musical sound, e.g. dynamic envelope, pitch contours, timbral development.
- Drawing something which reflects the emotional response to the sound, e.g. being "lifted" or "floating."
- A combination of the above.

Some of these alternatives are related to specific features present in either the sound itself or in the imagery of the sound-producing action, while others are more general in nature. This span from a detailed rendering of features to general sensations of the musical sound, is what could be called "variable acuity" [5].

A related experiment by the authors has been reported in [10]. Here, the sound-tracings have been extended to three-dimensional position data retrieved from a motion capture system, and are utilized for classification from a set of different sounds. Other experiments on feature extraction and classification related to musical movements can be found in [2]. However, this work is primarily undertaken on full body movements and the goal is to recognize emotional states.

The paper is organized as follows: The next section describes the data collection process, while Section 3 describes feature extraction from this data. Then Section 4 describes the setup of the experiments. Section 5 reports results from the experiments, while Section 6 discusses these results. Finally, Section 7 concludes the paper.

2. Data Collection

An A4 Wacom Intuos 2 graphical tablet was used for the study, giving information about XY position, pressure and XY rotation of the digital pen on the tablet.

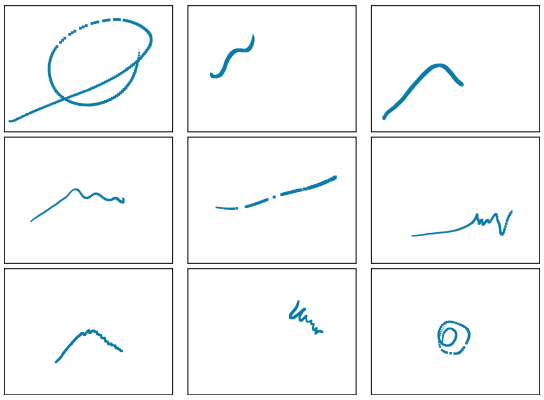


Fig. 1: Sound tracings for sound 5 from the 9 subjects.

Nine subjects were recruited ranging from novices to experts: 3 subjects with no or little musical training, 1 subject with little musical, but extensive dance training, and 5 subjects with music performance studies.

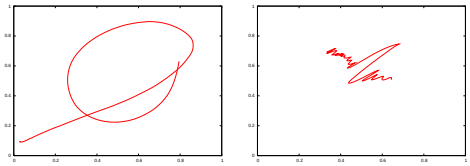
The subjects were presented with 30 short sounds and each of the sounds would be followed by a duration of silence equal in length to the sound being played. The instruction was to listen to the sound, and then draw the movement associated with the sound during the silence. We did not specify what type of drawings they should make nor which features in the sounds they should follow. For a sample of the results see Fig. 1.

The sounds were between 2 and 6 seconds in duration and were chosen to represent the three main sound types described by [11]: impulsive, sustained and iterative. The sounds were also selected so as to have different pitch and timbral/textural content: stable, changing/unstable or undefined. This combination of basic dynamical and pitch/timbre-related envelopes is the basis for a first and overall classification of sonic objects, what is called the *typology* in [11], and which may be supplemented by more detailed descriptions of internal features, e.g. various patterns of fluctuations, in what is called the *morphology* in [11].

3. Feature Extraction

From the raw data acquired as described in the previous section it is necessary to extract a set of features for input to a classifier system. From observation of the position plots of the tracings, it becomes apparent that the shape and position of the tracings can vary significantly between different tracings of a single sound, while some common qualities still can be perceived for many of the tracings. We therefore disregard positions and the overall shape of the tracings and concentrate on the features listed in Tab. I, which also includes example values from the shapes shown in Fig. 2.

The start and end points have been defined as the first sample with a positive pressure and the last sample with



(a) (b)

Fig. 2: Example tracings A (a) and B (b) from the same subject for two sounds, showing different qualities.

feature	Tracing A	Tracing B
duration	2.728	3.064
segment count	1	1
length	3.0	1.8
average speed	1.1	0.6
angle sum	16.5	62.5
curviness	6.1	20.4
growth	0.6	20.4
pressure sum	250.0	253.5
intensity	0.9	0.8
air distance	0.9	0.4
detour	3.1	5.0

TABLE I: Features and values for example tracings A and B, which can be seen in Fig. 2.

a positive pressure in the set of samples for one tracing. *Duration* is then the time difference between the start point and the end point. A tracing can be divided into several segments by having regions where the pressure equals zero, this defines the feature *segment count*. The *length* is the sum of distances between all consecutive sample points which have a positive pressure. The *average speed* is then simply *length* divided by *duration*. The calculation of *angle sum* consists of constructing two vectors from three consecutive points with positive pressure and calculating angle between them, and summing all the angles along the tracing. *Curviness* is then a normalization of this: *angle sum* divided by *duration*. The *growth* is the angle between the vector defined by the start point and the end point, and a horizontal line. *Pressure sum* is simply the sum of pressures for all sample points, while *intensity* normalizes this by dividing *pressure sum* with *duration*. *Air distance* is the shortest distance from the start point to the end point, not necessarily following the tracing, and *detour* the *length* divided by the *air distance*. All in all, these features resemble various types of qualitative features that can be observed from the sound-tracings, both dynamic and kinematic.

4. Classification experiments

With the developed set of features we have made an attempt at classifying the different sound-tracings into a given set of categories.

We have chosen to perform classification experiments on three different category subdivision schemes of the data set. The first subdivision scheme simply treats each

sound as a separate class, giving 30 classes in total. The two other schemes reduce the number of classes by grouping some of the sounds together according to the manual sound classification performed earlier. The most drastic scheme reduces the number of classes to three, by grouping the sounds by their main sound type, as described in Section 2. The final scheme reduces the number of classes to six, and is an attempt to have a more even distribution of classes to the feature vectors. Here, the subdivision is based on type, as in the three-class scheme, however some further subdivision based on pitch and timbre qualities is performed, as can be seen in Tab. II.

class #	sound type	internal qualities
1	impulsive	stable pitch/timbre
2	sustained	stable pitch/timbre
3	sustained	unstable pitch/timbre
4	sustained	stable pitch, unstable timbre
5	iterative	varying
6	impulsive	unstable pitch/timbre

TABLE II: Class description for the six-class scheme.

For classification, the support vector machine (SVM) method has been chosen [1]. SVMs employ a principle of structural risk minimization, which typically yields good generalization performance when compared to other classifier paradigms. For all category setup experiments C-SVMs with RBF kernels have been used. The parameters C and $gamma$ have been chosen separately for each of the category experiments through parameter search tools available in the LIBSVM package [4], which also do scaling of the input features. In addition to the SVM classifier, a verification experiment employing a reference classifier, k-nearest-neighbors (kNN), has been undertaken. A k value of 5 was used for all subdivision schemes. For evaluation of the classification accuracies, 10-fold cross validation with stratified sampling is performed. As an attempt to evaluate the relevance of the different features, feature weights are obtained through an application of the relief method [8], using 10 neighbors. The classification and feature weighting experiments have been conducted with the data mining framework RapidMiner [9].

5. Results

The classification results from the SVM classifier are shown in Tab. III. Accuracies are shown for the results on the training set, indicating the approximation performance, and the test set, showing the generalization performance. The application of the kNN classifier produced test accuracies of 17.8%, 40.7%, and 67.4%, for 30, 6, and 3 classes respectively. In order to get a more detailed insight into the classification result, the class recall percentages are shown in Figure 3. These numbers show the classifier's ability to detect the different categories.

	train	test
30 classes	49.6%	21.5%
6 classes	62.2%	44.4%
3 classes	75.6%	67.4%

TABLE III: Clssification accuracy with SVM.

6. Discussion

This section discusses the experimental results and indicates possible future improvements.

6.1. Classification Accuracy

The classification results from Table III indicate, as expected by visual inspection, that the classification of the sound tracings is a difficult task. It is however still uplifting that, even in the 30 categories case, it is possible to extract *some* information and perform classification which is considerably better than random guessing.

From the class recall results in Figure 3 one can see that in the case 30 categories subdivision scheme not all classes are detected, while some classes have a relatively high recall rate. This could be explained by several tracings looking relatively similar, such that either the different qualities of some sounds are not reflected in the tracings, or that the feature extractor is not able to distinguish such possible subtle features.

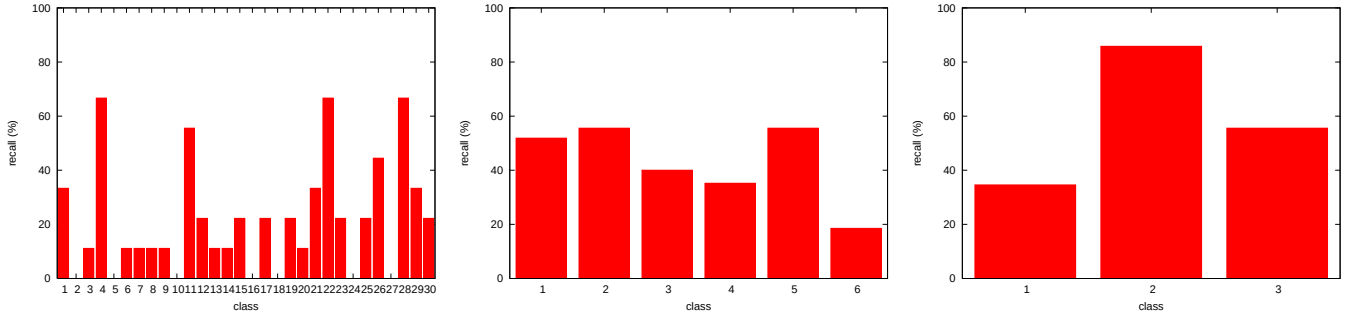
The three-class subdivision scheme gives relatively good classification results, however this is expected as the majority of the input vectors belong to class 2 (a classifier always predicting 2 would give 60% accuracy) and it is therefore more interesting to look at the six-class subdivision scheme, with a more equal distribution of input vectors over classes. Here, a relatively even class recall distribution is obtained. However, one can observe slightly worse performance with class 6, which is a subclass of the impulsive-type sounds with relatively few training samples. It is then interesting to note that class 5, which also has few training samples, but covers all of the iterative-type sounds, gives a good recall.

The classification results obtained from the kNN classifier were only slightly worse than the results from the SVM classifier, indicating that there is not very much to gain from the discriminative power of SVM over a simple approach like kNN on the current features.

It is also unclear how much better, if better at all, a human would be able to classify the tracings, given the big variations within each class and similarities between classes. Performing such an experiment would be interesting for further investigations.

6.2. Features and Data Set

It seems clear that the information extracted from the tracings is not sufficient for high precision classification,



(a) (b) (c)

Fig. 3: Class recall.

and we speculate that this is more due to the lack of information in the tracings themselves rather than a problem of the feature extraction.

Future improvements to the feature extraction could include the analysis of acceleration features, and possibly the acceleration sequence over time. In addition, it could be interesting to distinguish better between different types of ”curviness” such as ”jerkiness” which detects the amount of hard corners in the tracing.

Given the small number of tracings in the data set, the variations are large and it would be desirable with more data for the training. This could possibly reduce the number of ”noise” tracings, but depending on the subjects it could also open for even more interpretations giving more variety.

6.3. Suitability for Applications

It should be noted that the current classifiers have been trained on a data set with tracings from different subjects. For some applications, such as in navigation of sound collections, it could be possible to have one classifier per user. This would allow for an initial training phase where more consistent training data from the single user is collected.

It could be interesting for future applications to be able to extract information from longer, composite tracings. This could be achieved by local classification on extracted segments, or by treating the features as sequences which could be classified by hidden markov models or other temporal methods.

7. Conclusion

We have proposed a set of features for classification of music-related actions in the form of sound-tracings, and analyzed the performance of classification based on these features. Although higher classification accuracy is desired, it is shown that some useful information can be extracted from subjective sound-tracings. Factors which could improve classification accuracy include: a more extensive data set, the training of classifiers for single users, and further extension and refinement of the feature extraction. The abovementioned points are subject of future work, as well as performing a study on human classification performance in order to have a basis for comparison.

References

[1] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[2] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. Multimodal analysis of expressive gesture in music and dance performances. *Lecture notes in computer science*, 2915:20–39, 2004.

[3] C. Casciato, A. R. Jensenius, and M. M. Wanderley. Studying free dance movement to music. In *Proceedings of ESCOM 2005 Performance Matters! Conference*, Porto, Portugal, 2005.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing. - a preliminary study. In *2nd ConGAS International Symposium on Gesture Interfaces for Multimedia Systems, May 9-10 2006*, Leeds, UK, 2006.

[6] R. I. Godøy, E. Haga, and A. R. Jensenius. Playing air instruments: Mimicry of sound-producing gestures by novices and experts. In S. Gibet, N. Courty, and J.-F. Kamp, editors, *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop, GW 2005, Berder Island, France, May 18-20, 2005, Revised Selected Papers*, volume 3881/2006, pages 256–267. Berlin Heidelberg: Springer-Verlag, 2006.

[7] R. I. Godøy, A. R. Jensenius, and K. Nymoen. Chunking in music by coarticulation. *Acta Acoustica united with Acoustica*, 96(4):690–700, July/August 2010.

[8] K. Kira and L. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, pages 129–129. JOHN WILEY & SONS LTD, 1992.

[9] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. Yale: Rapid prototyping for complex data mining tasks. In L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM.

[10] K. Nymoen, K. Glette, S. Skogstad, J. Torresen, and A. Jensenius. Searching for Cross-Individual Relationships between Sound and Movement Features using an SVM Classifier. In *Proceedings of the 2010 Conference on New Interfaces for Musical Expression (NIME)*, 2010.

[11] P. Schaeffer. *Traité des objets musicaux*. Paris: Editions du Seuil, 1966.