Sound Selection by Gestures

Baptiste Caramiaux, Frédéric Bevilacqua, Norbert Schnell

Sound Selection by Gestures

Baptiste Caramiaux
IMTR Team
Ircam - CNRS
Paris, France
caramiau@ircam.fr

Frédéric Bevilacqua
IMTR Team
Ircam - CNRS
Paris, France
bevilacq@ircam.fr

Norbert Schnell
IMTR Team
Ircam - CNRS
Paris, France
schnell@ircam.fr

ABSTRACT

This paper presents a prototypical tool for sound selection driven by users' gestures. Sound selection by gestures is a particular case of "query by content" in multimedia databases. Gesture-to-Sound matching is based on computing the similarity between both gesture and sound parameters' temporal evolution. The tool presents three algorithms for matching gesture query to sound target. The system leads to several applications in sound design, virtual instrument design and interactive installation.

Keywords

Query by Gesture, Time Series Analysis, Sonic Interaction

1. INTRODUCTION

The study presented in this paper is part of a series of studies concerning the analysis of the relationships between movements and sounds for the design of virtual instruments and more generally for applications in sonic interaction. Consider the following scenario. A user imagines a sound that is too abstract to be described using words. Possibly, a skilled user should be able to sketch with the voice what the sound looks like. Here we consider the case where the person uses gestures. If the profiles drawn by the temporal evolution of the sound's characteristics is clear in the users' mind, they could try to gesturally "trace the sound" either in the air or on a surface. Thus, the goal of the proposed tool is to return a sound that is the most pertinent according to the tracing of the performed gesture. The problem is a particular case of "query by content" in multimedia databases. The input gesture is usually called the *query* and the resulting sound the *target*.

1.1 Background

The general problem of "query by content" in multimedia database was extensively studied and the literature is flourishing. In Music Information Retrieval (MIR), a particular case of "query by content" is the famous "query by humming" problem [4]. Query by humming system allows the user to find a song in a database by humming part of the tune. Most researches into query by humming use the notion of *contours* that is the the sequence of relative differences in pitch between successive notes. Another illustrative example is the "query by tapping" system [6] that allows the user to find a song by tapping the rhythm. This system is based on onset detection and temporal alignment.

On the gestural counterpart, there is a dramatic lack of literature about audio query by gesture systems in either the NIME or MIR communities. Previous works are more dealing with the inverting system that is analyzing which gesture is performed by a user while listening to a sound [5]. When trying to match a gesture and a sound, two problems occur: which features should we select for describing either the gesture or the sound? how can we fill the informational resolution gap between both signals?

1.2 Proposed system

Figure 1 illustrates the system for gesture-driven sound selection. A user performs a gesture that matches, at least from the user perspective, an abstract sound. After a preprocessing module, the system contains several algorithms for time series multimodal matching. Each algorithm retrieves a specific part of information in the relationship between gesture and sound. The algorithm is the choice of the user. The matching algorithm returns the sound index in the database together with a score that indicates the target pertinency. Finally the sound is played to the user.

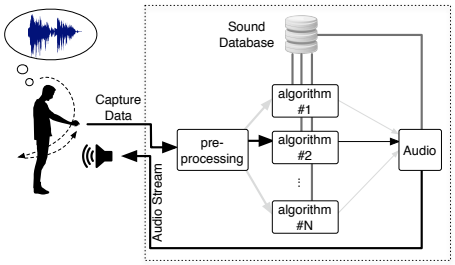

Figure 1: The proposed system. A sound that best matches the input gesture is found in a database. The matching depends on the algorithm used.

2. PROTOTYPE

In this section, we present the implementation. The algorithms used in the current version are reported in the next section. Then the available implementation in the Max/MSP software is described.

2.1 Algorithms

Each matching method allows for retrieving specific information in the relationship between gesture and sound.

Correlation-based selection

The method is based on the correlation between the input gesture parameters and the sound features [3]. The method

NIME'11, 30 May–1 June 2011, Oslo, Norway.

is similar to Principal Component Analysis but adapted for two datasets of different dimensionality. It is called Canonical Correlation Analysis (CCA). The algorithm finds the principal (or canonical) components that explain the most the covariance between the two datasets. Then, it returns two new sets of variables (for both gesture and sound) that are ordered from the most correlated (the first ones) to the less correlated (the last ones). Sound selection tool based on CCA allows for the selection of the predominant features (in terms of correlation) from both gesture and sound parameters. The first correlation coefficient (i.e the maximum) is used as the similarity score. A sound is selected if the variation of a combination of its features is similar to the variation of a combination of the gesture parameters. Since correlation is computed sample-by-sample, a high score also indicates that gesture is synchronous to the sound. Finally the sound is selected at the end of the gesture leading to the need to mark the beginning and the end (e.g. using a button).

Time-warping based selection

One can ask to preserve the inherent variability in gesture and choose as similarity criterion the global shape matching and the coherence between amplitudes. To that extent, the second strategy is based on temporal alignment of both multidimensional signals. The method returns a score that depends on whether the two signals are far from each other in terms of alignment and amplitudes. This method is an HMM-based technique that has been used for gesture recognition and following [2]. It is computationally efficient, multidimensional, real time and makes use of a simplified learning process. For sound selection tool, a sound is selected if the user performs a gesture that evolves similarly to the sound features but can be non-linearly time shifted. Here real time means that a sound is selected while the gesture is performing. However, it requires to previously select the features chosen to be matched.

An hybrid strategy

The algorithm is iterative and uses both correlation-based measure and temporal alignment. The strategy is to compute CCA between the user's gesture taken as input and all the sounds in the database. Then, we take the sound corresponding to the highest correlation coefficient and we apply a temporal alignment between the projected correlated variables. We then iterate using the aligned gesture and the original sounds in the database. The use of temporal alignment is two-fold. First it allows to better discriminate the candidate sound from the other. Second, it allows to precise which feature is actually predominant in the mapping user's gesture-to-selected sound. The iterative process is heuristic but results to always increase the correlation coefficient. Using this strategy allows for more temporal flexibility without constraining the system by fixing previously the features but is computationally time-consuming.

2.2 Implementation

The various algorithms are encapsulated in an application developed in the Max/MSP real-time programming environment (and MnM [1]). The sound pool uses MuBu [7] that contains N sounds together with their audio descriptors. These audio descriptors are directly computed in Max/MSP. The motion data are received by OSC allowing for the use of a wide range of interfaces. When the analysis is done, the program returns the index of the best matching sound belonging into the database, and it is visualized in the MuBu editor (see figure 2 for a screenshot of the tool).

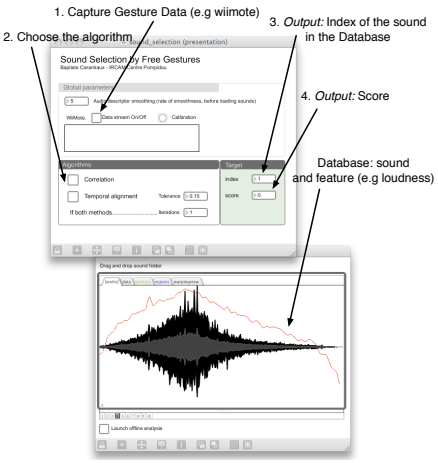

Figure 2: The Max/MSP patch for sound selection by free gestures. The example is given with one feature (loudness) per sound and is used with a WiiMote controller. The user can choose which algorithm is used for the time series matching.

3. CONCLUSION

In this paper, we presented an application allowing for sound selection driven by user's gestures. The application computes the similarity between the gesture and sound parameters' temporal evolution. The tool aims to embed several algorithms for time series matching. A version has been developed in the Max/MSP software and uses MnM.

Finally, we have recently investigated by an experimental study how people associate gestures to environmental sounds for which either the cause having produced the sound can be identified or not. This study will give important insights for the relationships between gesture and sound and will help for the design of new algorithms.

4. ACKNOWLEDGMENTS

We acknowledge partial support from the project Interlude -ANR -08-CORD-010 (French National Research Agency).

5. REFERENCES

[1] F. Bevilacqua, R. Müller, and N. Schnell. Mnm: a max/msp mapping toolbox. In *Proceedings of the 2005 conference on NIME*, pages 85–88. National University of Singapore, 2005.

[2] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. In *Lecture Notes in Computer Science (LNCS)*. Springer Verlag, 2009.

[3] B. Caramiaux, F. Bevilacqua, and N. Schnell. Towards a gesture-sound cross-modal analysis. *Lectures Notes in Computer Science, Springer-Verlag*, 2009.

[4] R. Dannenberg, W. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis. A comparative evaluation of search techniques for query-by-humming using the musart testbed. *Journal of the American Society for Information Science and Technology*, 58(5):687–701, 2007.

[5] R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing: A preliminary study. In *Proceedings of the COST287-ConGAS 2nd International Symposium on Gesture Interfaces for Multimedia Systems (GIMS2006)*, 2006.

[6] J. Jang, H. Lee, and C. Yeh. Query by tapping: A new paradigm for content-based music retrieval from acoustic input. *Advances in Multimedia Information Processing*, pages 590–597, 2001.

[7] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, R. Borghesi, et al. Mubu & friends-assembling tools for content based real-time interactive audio processing in max/msp. In *Proceedings of the ICMC, Montreal*. Citeseer, 2009.