Study on Gesture-Sound Similarity

Baptiste Caramiaux, Frédéric Bevilacqua and Norbert Schnell

IRCAM, CNRS-UMR STMS, 1 Place Igor Stravinsky, 75004 PARIS, France
baptiste.caramiaux@ircam.fr
frederic.bevilacqua@ircam.fr
norbert.schnell@ircam.fr

Methodology

The dominant paradigm in the design of gestural sound control is based on mapping gesture parameters to sound synthesis parameters. The main difficulty resides in choosing an appropriate mapping strategy between low or high level parameters [4]. Our long term goal is to propose adaptive methods to set such a mapping automatically from gestures performed while listening to a sound. Such methods require a co-analysis of the performed gesture and the listened sound.

This paper focuses on notions of similarity between gesture signals and audio signals. Quantitative analysis of gesture data together with audio data were proposed mostly based on second-order moments as correlation applied between sensor values and sound description parameters (see [2], [5]). These methods permit to highlight how a human can synchronize with the tempo and allows for the determination of the most important features in gesture in relationship to the sound. However, these methods suffer from important limitations since the relationships between sound and gesture are considered as linear or instantaneous.

To overcome such shortcomings, this study proposes to define similarity measure as a higher-order statistical measure between data. Inspired by previous works [3] in the Information Retrieval (MIR) field, we propose to use Mutual Information (MI) between signals. Actually, similarity measures will not be computed on the signal itself but on its probability distribution encoding higher level aspects as degree of prediction. Inspired by this literature, we are interested in the extraction of the information content of gesture and sound.

Practically, the first step in the modeling is to estimate the probability distribution function of each signal. The gesture signal is typically the temporal evolution of kinematic variables and we use audio descriptors computed on a sliding window to describe the sound. The second step is the use of an appropriate divergence (e.g. Kullback-Leibler divergence) to measure how much one signal can be explained by the other. If the divergence returns zero or lower than a minimal ϵ-value (corresponding to an objective criteria), then the signals are assumed to be similar. On the contrary to correlation-based methods, here the gesture is no more constrained to be synchronized with the audio signal.

Results and Discussion

In this section we present experimental results obtained applying this methodology on real data. These data has been collected in May 2008 in the University of Music in Graz. For the experiment 20 subjects were invited to perform gestures while listening to a sequence of 18 different recorded sound excerpts. The sound corpus included a wide variety of sounds. The gestures were performed with a small hand-held device that included markers for a camera-based motion capture system recording its position in Cartesian coordinates.

We consider sampled overlapping windows of the gesture signal and the audio descriptor signal, $\boldsymbol{G}_i$ and $\boldsymbol{D}_i$ where i is the time index (in samples) since the beginning. Each windowed signal is a stochastic process of N random variables. We compute their variance on the window: $\left(\sigma_1^k, ..., \sigma_N^k\right)$ where k is the frame number. The resulting variance vector $\boldsymbol{\sigma}^k$ is considered as a set of independent and identically-distributed random variables on which we estimate the density function. Since we estimate the variance signals as Gaussian mixtures, we choose the KL-divergence to measure the similarity between the two probability density functions (see [1]).

Using the aforementioned data, we analyze the gestures performed while listening to three different sounds: a sound of an ocean wave; a solo flute (from *Sequenza I* for Flute by L. Berio); a sound of a crow's caw. These sounds correspond to three distinct morphologies. In conclusion, we find that the KL-divergence is minimized when we consider the gesture performed on one sound and the corresponding audio descriptor whereas its value is higher for a mixed case including a gesture or an audio descriptor from other sounds. Lastly, we compute the KL-divergence considering the audio descriptor of a sound (e.g. the wave) and the corresponding gesture performed by each candidate. This gives a similarity evaluation of each performance that we can compare with videos in order to discuss its pertinency.

References

1. Banerjee, Arindam, Merugu, Srujana, Dhillon, Inderjit S., and Ghosh, Joydeep. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.
2. Caramiaux, Baptiste, Bevilacqua, Frédéric, and Schnell, Norbert. Towards a gesture-sound cross-modal analysis. In *Lecture Notes in Computer Science*. Springer Verlag, 2009.
3. Foote, Jonathan. A similarity measure for automatic audio classification. In *Proceedings AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Copora*, American Association for Artificial Intelligence, 1997.
4. Hunt, Andy and Wanderley, Marcelo M. Mapping performer parameters to synthesis engines. *Organised Sound*, 7(2):97–108, 2002.
5. Luck, Geoff and Toiviainen, Petri. Ensemble musicians synchronization with conductors gestures: An automated feature-extraction analysis. *Music Perception*, 24(2):189–200, 2006.