Report - Kansei the Technology of Emotion Workshop
Rapport de Mission

Marcelo M. Wanderley
IRCAM - Groupe Analyse/Synthèse

January/98

Contents

Abstract

This report intends to comment on some of the papers presented at the KANSEI - the technology of Emotion workshop, held on the 3rd and 4th October/97 in Genova - Italy.

Chapter 1

Introduction - KANSEI - The Technology of Emotion Workshop

Organized by the AIMI - Italian Association of Musical Informatics, the Musical Informatics Laboratory of the University of Genova (DIST) and chaired by Antonio Camurri, president of AIMI.

And what does the word "KANSEI" mean?

Quoting A. Camurri's presentation of the workshop:

> Kansei is a Japanese word referring to "emotion", in the sense of acquired sensibility towards art and music as a whole. "Kansei information processing" or "technology of emotion" refers to methodology, tools and applications related to emotion modelling, to analysing components of human behavior which we call *emotional*, with respect to the traditional research in artificial intelligence concerning *rational* behavior.

The workshop focused mainly on the technology of emotion in music. The main topics covered were:

- *expressivity in virtual music instruments*,
- *adaptative hyper-instruments*,
- *computational models of artificial emotions*,
- *relations between music and gesture languages*,
- *on-stage real-time multimodal environments*,
- *interactive dance/music systems*, and
- *modelling of expressive performance*, among others.

Chapter 2

Comments on selected papers

We will provide here some comments on selected papers on expressivity and gestural capture/control of synthesis and also try to relate these works to previous and concurrent research on these areas.

2.1 *S. Hashimoto and the research about KANSEI in Japan*

S. Hashimoto classifies information technology in three categories:

- physical signal processing,
- semantic symbol processing and
- emotional (KANSEI) information processing.

He considers that the technology for art belongs to the third item.

He further classifies these three categories according to their differences in evaluation. As an example: for signal processing, evaluations is by means of *measurement*; for artificial intelligence, *recognition*; for kansei technology, *appreciation*.

He considers that a system to create art with human (sic) has to understand not only the user's intentions, but also the environment of performance with multi-modal sensing ability. The system will try to understand the will of a performer while iterating bi-directional communication.

The author considers the most important emotional information in human gestures as the forces applied by the body (that can be computed directly from acceleration of body movements). He opposes this approach to the most general one, consisting of recognizing shapes and positions (of the body). His group at Waseda University has therefore developed new musical environments driven by gestures and employing acceleration sensors and data-gloves as gesture input devices. They also make use of neural networks

(NN) on the processing of gesture information stage, since using NN one does not need knowledge on the Kansei processing itself.

Part of this work can be found in [1] and [2].

Finally, he cites some facts about the current Kansei research in Japan. According to the author, in 92, the Japanese Ministry of Education started a three year grant program in "Kansei information processing", which was followed by another special grant program on "Virtual Reality". Over 50 research groups from different fields (computer science, physiology and psychology) have joined the program and were divided in four groups: Kansei modelling, kansei information in media, kansei in human behavior and kansei in communication.

2.2 *A. Camurri and the research on gestural capture, dance/music interfaces and Artificial Intelligence at DIST, Genova - Italy*

A. Camurri and his group at University of Genoa have been developing research on gestures for more than 10 years, usually on the area of dance-music interfaces [3] and artificial intelligence applied to music [4]. His group presented several articles at the workshop. We will focus here mostly on the paper: Towards KANSEI information processing in Music/Dance interactive multimodal environments (page 74 of the proceedings). Abstracts of other papers can be found through the GDGM home-page.

In this paper the authors' focus is on movement and gesture analysis and machine communication to humans. In particular, on the first topic, the interest is mainly on recognition of non-symbolic, expressive data from human movement and gesture and their relation to music performance.

The first part of the article deals with the general framework of this work: *human-computer communication in interactive Multimodal Environments (ME)*. The authors consider an ME as an active space populated by agents allowing one or more users to communicate by means of different modalities, such as full-body movement, gesture, voice. Users get feedback from the ME in real-time in terms of sound, music, visual media, lights control, and actuators (e.g. mobile scenography, on-stage robots). The interest in kansei information processing derives from the fact that high-level interaction requires, according to the authors, ME agents to be capable of changing their "character" and social interaction over time.

The article follows by discussing human gesture taxonomies, mainly by Coutaz [5] and Cadoz [6], kansei and movement analysis, where the authors are concerned about extracting high-level, whole body features, gesture gestalts, from the observation of a dancer or a performer.

Systems used for movement and gesture detection include camera based ones, where a preprocessing phase tries to recognise the posture of the human figure and match it to one of a few stereotypical postures or clusters, by extracting a small number of parameters and by sending them as inputs for a self-organizing neural network. The second phase consists of a possible segmentation on the stream of frames and application of different analysis algorithms to extract features from the different video segments. Other systems are the *V-Scope*, that uses ultrasound and infrared technologies, developed by Lipman Ltd. and *DressWare*, wearable piezo-resistive fabrics, developed by De Rossi et al. [7].

The data acquisition system used is called *DanceWeb* and is able to handle simultaneously up to 16 digital inputs and 40 ultrasonic sensors, configured in 8 groups of 5 sensors, each sensor individually enabled. Sample time can be set via software and varies between 15ms and 10 s. The system communicates via an RS 232 to Win32 compatible applications.

Finally, data processing is done by a software developed at DIST, called *Mummia*, an environment for the development and supervision of real-time, dance- and gesture-driven performances. Material provided by the composer is dynamically rearranged according to composer's rules and external inputs, typically dancers' movements. "Mummia" is based on two notions: *virtual sensors* and *MIDI agents*. A *virtual sensor* can be: a physical sensor or an elaboration of it (e.g., its derivative), or the fusion of data from different sensors. A *MIDI agent* is an agent able to produce a MIDI output according to its internal state, to continuous parameters and triggers that modify its behavior (typically, a single agent only controls a small portion of the global score/performance, for a given duration of time. Conversely, more than one agent can control a single musical parameter).

The second part of the paper deals with the possibilities to control the movement of physical objects, including robots and in general effectors. The authors discuss the robot control system and briefly the development of a robot's emotional component. These developments have been implemented in the project called "La Cita dei Bambini", a permanent exposition in Genova.

More information on Camurri's previous and related work can be found at [8], [3], [4], [9], [10] and [11].

2.3 *I. Zannos and P. Modler and gestural capture at the Staatliches Institut für Musikforschung, Berlin - Germany*

In the first of the two papers presented at the workshop, the authors describe ongoing research on a real-time system for gesture controlled music performance. Two frameworks are compared:

The first one is developed by the authors and uses the "JET" - Java Environment for TEMA, connected to a prototype of the MIDAS system for real-time open distributed processing (University of York). The idea here is to bring advantages of open heterogeneous networks to experimental music design and to performance systems.

The JET architecture is based on the MIDAS signal processing engine, with low level drivers (written in C), input and musical data-handling processes (Java), a mid level GUI (also in Java) and a high-level graphical data representation (written in Java and/or VRML). Java processes take over the tasks of GUI, user interaction management and also high-level data management. They send configuration and performance instructions to MIDAS, which produces sound and/or animated processes. Both (Java and MIDAS processes) can be distributed over the network. Once sound processing on MIDAS is configured, it can be controlled via MIDI, process serial input from the *SensorGlove* (see below), or mouse and a computer keyboard.

The second framework is based on the program SuperCollider for Apple Macintosh and is used in order to identify which high-level programming features are required. This framework will later be implemented in JET by the authors, since the program is only available for Macintosh. The authors use it as a working paradigm for programmable real-time sound processing architecture.

Both frameworks represent gesture data input from a data-glove as well as data for performance control and sound synthesis configuration. According to the authors, the three main tasks of this work are (1) provide a flexible prototyping environment so that experiments with different approaches can be designed, (2) devise mappings of the input parameters to sound production such that the relationship between hand movements and the resulting sound is intuitively graspable for the user, but not trivial, and (3) apply features and gesture recognition techniques to achieve high-level communication with the computer.

The systems are controlled by the *SensorGlove*, a custom data-glove constructed at the Technical University of Berlin, used as a multi-parametric input device for music. The *SensorGlove* is a custom gesture input device that has sensors measuring the bending angle for every joint of each finger (providing a resolution less than 1/10 of a degree in a finger joint flexion measurement) as well as hand movement acceleration. The transmission sampling rates are up to 200 Hz and the resolutions are between 7 and 14 bits.

Gesture data processing involves three steps in this system: preprocessing (data range calibration and scaling, preliminary feature extraction), gesture recognition and mapping to or generation of performance parameters.

To perform gesture acquisition, preprocessing is done in order to improve the quality of the input by filtering noise, smoothing rapid jumps and detecting and undoing wrapping of the input values.

The GUI consists of a central panel with buttons for operating glove routines and for recording, storing and replaying gestures, numeric and slider display of the input parameters and switches for routing the parameters to adjustable MIDI parameters and also a graphic display of the input parameters.

2.3.1 *Mapping:*

The mapping can consist of simple mapping of the input parameters to synthesis parameters or/and feature extraction with Time Delay Neural Networks (TDNN)[1], in order to combine direct control of sound with higher level semantic processing of gestures.

In the second paper the authors report that using only direct mapping of single sensor values (12 finger ankles and 3 hand acceleration values) to sound parameters can provide good results concerning the possibilities of controlling parameters of the sound algorithm (FM, granular, analog synthesis), but that coordinated control of a larger number of connected sensors was difficult. This effect seems natural since coordinated control of many parameters usually requires a leaning phase, and they have reported that the performer had a short time to get familiar to sensor functions. Their conclusion was that conscious control of a large number of parameters is not possible with such directly connected sensors (i.e. one-to-one mapping).

The next control (mapping) strategy tested was the use of ANNs (artificial neural networks) that have been trained with the gestures of the sensor glove. The glove is connected to a Macintosh and the sound generation is performed by synthesizers, SuperCollider or audio MAX/ISPW (FTS), controlled via MIDI.

2.3.2 *Gesture recognition:*

Three hand gestures have been used and the net was trained by a set of 8 gesture recordings for each pattern — 2 sets each of 4 recordings with different timing levels: very slow, slow, fast, very fast.

They report "good" recognition results for the gestures presented (with different velocities). Gestures are presented in real-time and at each sample time, the net gives back an output vector, indicating the recognition of the pattern. They noticed that the net responds on moving phases of the gestures more than on stable phases.

The authors further propose a distinction between possible hand gestures in a Symbolic level and a Parametric level. A separation of gestures in sub-gestures as: (1) five finger gestures - joint angles and spreading, and (2) one hand gesture - translation and rotation of the whole hand. They consider that with the sub-gesture architecture, gesture sequences can be recognized which use only a part of the hand as a symbolic sign, while other

[1] SNNS - Stutgart Neural Network Simulator, at: http://www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html

parts of the hand movements are used as parametric signs. As an example, they propose a straightened finger indicating mouse down (symbolic gesture) and moving the whole hand determining the changing of the position of the mouse (parametric gesture). This way, different sub-gestures (symbolic or parametric) can be mapped in different ways to sound parameters.

2.3.3 *Example: Hit detection using sub-gestures.*

The net was trained to detect hits of a single finger, using derivates of higher order to indicate the hit — the second derivative of the finger sub-gesture is proportional to the occurrence of the force and the third derivative to the changing of the force.

They finish the paper by discussing topics relating interactive computer systems, emotions and neural nets.

More information can be found in [12].

2.4 *A. Mulder and the gestural research in ATR, Kyoto - Japan*

According to the authors, the goal of this research is to develop gestural interfaces that allow for simultaneous multidimensional control, such as in musical composition and sound design - control of timbre, involving the control of many inter-dependent parameters simultaneously. They claim that, in order to reduce the cognitive load when performing simultaneous multidimensional control, it is necessary to design a human computer interface that implements data reduction and/or an interface that exploits the capability of human gestures to effortlessly vary many degrees of freedom simultaneously. They also consider that, while the human hand is well suited for multidimensional control, due to its detailed articulation, most general interfaces do not exploit this capability due to a lack of understanding of the way humans produce their gestures and what meaning can be inferred from these gestures.

Their approach consists of focusing on the continuous changing represented by the gestures produced by the user, as opposed to (1) recognition of gesture formalisms (need for learning the formalism) and to (2) natural gestures recognition, since they report previous results where the latter is considered by the authors as rarely sufficiently accurate due to classification errors and segmentation ambiguity. They also consider that touch and force feedback can be replaced by only acoustic feedback (with some compromises — not specified). This option was chosen due to technical constrains in implementing touch and force feedback in a shape manipulation task.

The proposed system uses MAX/FTS running on a R10000 SGI Onyx workstation with audio/serial option to interface two Virtual Technologies

Cybergloves and a Polhemus Fastrak sensor. Some considerations are made in respect of accuracy of the glove and calibration, considered tedious, as well as specific to each individual user. The authors have developed new FTS objects specific for facilitating quick and easy prototyping of various gestural analysis computations and allowing for application of the computations to different body parts.

They consider both sound and human movement as able to be represented at various abstraction levels and claim that a mapping will be faster to learn when movement features are mapped to sound features of the same abstraction level. The strategy is to use a shape as a means to relate hand movements to sound variations.

The system works exploring the intuitive relations between shape of physical objects and timbre, as well as shape and manipulation for the design of a sound editing environment, where the user can change the sound by applying shape orientations to a virtual object. Shape features are subsequently computed and mapped to sound parameters.

2.5 *I. Choi and R. Bargar, HCI at the University of Illinois - USA*

2.5.1 *Interactivity vs. Control*

After some discussion on the experience gained in previous works and on philosophical reasoning on machine attributes in human-machine performance, I. Choi discusses the notion of Interactivity versus Control. She begins by specifying the notion of interactivity in terms of system connectivity and parametrization.

In the system described, movements effect system states defined as machine responses. Such responses are registered as measurements taken at an array of pressure sensors - in this case, sensors placed at specially designed shoes. Other input devices include magnetic tracking, computer vision and positional devices. These registered measurements undergo an interpretation by a fuzzy inference process, noting that only the measurements in the range of state changes will be reflected in the machine response.

The inference process then returns an output value which is passed to response dynamics encoded in the machine, representing a range of motion in the computational model. The motion then determines the change of states in effector functions that describe output signals (such as graphics or sound).

The author then discusses the notion of parametrization of interactive parameters. As an example, parametrization of a brass tone synthesis algorithm can be made in terms of pitch, loudness and spectral distribution of sinusoidal components, etc. or, in the case of physical models, breath pressure, lip pressure, etc. Choi considers the choice of parametrization as capable of trivializing or enhancing interactivity in a human-machine relationship.

Next the author discusses the relationship between the observers of a musical performance and the musician/performer movements (as present in musical experience) as a perceptual interface to the process of listening. She considers that in the proposed system, the performer also plays a role of an observer, and that his/her actions are seen by other observers both (traditionally) as part of the musical performance and (non-traditionally) as an expertise accessible to other observers, as opposed to virtuoso music performance conditions. She considers that this accessibility changes the mode of other observers' spectation.

2.5.2 *The manifold interface and the manifold controller (MC):*

The manifold interface provides an interface to parameter control spaces with more than three dimensions by means of graphical lines and surfaces. It allows the user to navigate in a high-dimensional parameter space from a visual display with continuous gesture input (with at least two degrees of freedom).

The manifold controller (MC) is a set of classes linking graphics, hardware input devices, and sound synthesis engines. It can be defined as an interactive graphical sound generation tool and composition interface. Computational models involve sound synthesis models (also physically-based systems) and composition algorithms.

The idea behind the manifold interface is that of organizing control parameters in order to provide efficient system access. They also seek to provide a representation of the system that has visual simplicity. This is done by proposing the concepts of control space, phase space and window space (please refer to [13] for details).

Control space refers to both phase space and window space as a whole. Phase space is an n-dimensional space where vector arrays correspond to states of a parametrized system. It represents all the permissible combinations of parameter values of an algorithm.

In order to make it possible to visualize the n-dimensional phase space, the authors [13] propose the concept of a window space, that represents the mapping of the phase space data to a three dimensional space. Said in another way, it defines how a three-dimensional visual representation is embedded in the high-dimensional space. Therefore, an observer may control the window space by panning and zooming in the phase space. Citing the author: "The window space provides a domain for generating and modifying classes of control point sets. These points represent combinations of parameter values as user-specified, and they are associated with particular sounds".

This association is reported to enhance the ability to identify boundaries where character shifts occur in the states of the system.

Choi continues by considering two main gestural acquisition system types: externalized systems, where a sensing device makes an observational measurement according to world-centered coordinates and human-centered systems, where the measurements share the coordinate system with the observer. She claims that the information cues sent by human (body)-centered systems are complementary to world-centered positional cues (for example, from the performer's eyes) and to gravity-centered balance cues (from the inner ear).

She has implemented an example of human-centered system by designing a foot-mounted interface, sensitive to natural stance and bipedal locomotion. Four pressure sensors are placed on the base of the boot at the heel, the inner and outer foot ball and at the toe tip. The outputs of these 8 sensors (2 feet) are then sent to an inference system based on fuzzy rules. The result of the fuzzy process will then control both graphic engines and sound synthesis engines in real-time.

2.6 *Gestural Research at IRCAM*

For a survey on current gestural research projects at IRCAM, the reader is directed to two main sources - the home pages of the "Groupe de Discussion sur le Geste Musical".

The internal page[2] contains information related to talks and events taking place at IRCAM and around Paris and also different resources about sensors and musical pieces/gesture related projects at IRCAM's Pedagogy Department.

The second page[3] is a public one (as opposed to the first one, accessible only from inside IRCAM) and is more directed to a discussion on different subjects related to gestural capture control. It provides, as well as links to projects inside IRCAM, links to researchers in different countries and their projects.

2.7 *Conclusion*

We would like to emphasize the works commented here and presented by 6 different groups, in Tokyo and Kyoto, Genova, Berlin, Illinois and at IRCAM. These papers show the high interest in the area of gesture capture/recognition and gestural control.

Nevertheless, many other interesting papers were also presented and the reader is directed to the GDGM page for abstracts of some of these papers,

[2] at: *http://www.ircam.fr/equipes/analyse-synthese/wanderle/Gestes/Interne/index.html*

[3] at: *http://www.ircam.fr/equipes/analyse-synthese/wanderle/Gestes/Externe/index.html*

or to the proceedings of the workshop, available at the Analysis/Synthesis group.

The workshop also proved very useful for contacts with other researchers, mainly A. Camurri, A. Mulder, P. Modler and D. Arfib. As one of the results, the external gesture page from the GDGM was created and a collaboration with Camurri is envisaged.

Our participation was made possible due to financial support for the tickets and workshop inscription taxes graciously provided by the Analysis-Synthesis Group, support we would like to acknowledge here.

Paris, January 26th 1998.

Marcelo M. Wanderley

Marcelo.Wanderley@ircam.fr

2.8 *Aknowledgements*

Thanks to Diemo Schwarz for reading and commenting this manuscript.

Bibliography

[1] H. Sawada, N. Onoe, and S. Hashimoto, "Acceleration sensor as an imput device for musical environment," in *Proc. Int. Computer Music Conf. (ICMC'96)*, pp. 421–424, 1996.

[2] H. Sawada, N. Onoe, and S. Hashimoto, "Sounds in hands - a sound modifier using datagloves and twiddle interface," in *Proc. Int. Computer Music Conf. (ICMC'97)*, pp. 309–312.

[3] A. Camurri, "Interactive dance/music systems," in *Proc. Int. Computer Music Conf. (ICMC'95)*, pp. 245–252, 1995.

[4] A. Camurri, A. Catorcini, C. Innocenti, and A. Massari, "Music and multimedia knowledge representation and reasoning: The harp system," *Computer Music J.*, vol. 19, no. 2, pp. 34–58, 1995.

[5] J. Coutaz and J. Crowley, "Interpreting human gesture with computer vision," in *Proc. Conf. on Human Factors in Computing Systems (CHI'95)*, 1995.

[6] C. Cadoz, "Le geste canal de communication homme-machine. la communication «instrumentale»," *Sciences Informatiques - Numéro Spécial: Interface Homme-Machine*, vol. 13, no. 1, pp. 31–61, 1994.

[7] D. Derossi, A. Dellasanta, A. Mazzoldi, and V. Tinucci, "Wearable piezoresistive fabrics for monitoring human body kinematics," in *Proc. KANSEI - The Technology of Emotion Workshop*, pp. 98–100, 1997.

[8] A. Camurri, *Applications of Artificial Intelligence methods and tools for music description and processing*, vol. 9 of *The Computer Music and Digital Audio Series*, pp. 233–266. A-R Editions, 1992.

[9] A. Camurri, A. Coglio, P. Coletta, and C. Massucco, "An architecture for multimodal environment agents," in *Proc. KANSEI - The Technology of Emotion Workshop*, pp. 48–53, 1997.

[10] A. Camurri and P. Ferrentino, "A computational model of artificial emotion," in *Proc. KANSEI - The Technology of Emotion Workshop*, 1997.

[11] A. Camurri and M. Leman, *Gestalt-Based Composition and Performance in Multimodal Environments*, vol. Music, Gestalt and Computing of *Studies in Cognitive and Systematic Musicology*, ch. V, pp. 495–508. Springer Verlag, 1997.

[12] P. Modler, *Interactive Computer Systems and Concepts of Gestatlt*, vol. Music, Gestalt and Computing of *Studies in Cognitive and Systematic Musicology*, ch. V, pp. 469–481. Springer Verlag, 1997.

[13] I. Choi, R. Bargar, and C. Goudeseune, “A manifold interface for a high dimensional control space,” in *Proc. Int. Computer Music Conf. (ICMC'95)*, pp. 385–392, 1995.