Musical Interface to Audiovisual Corpora of Arbitrary Instruments

Joachim Gossmann
1 Springfield Rd.
Bangor, Co Down
N Ireland, BT20 5BZ
joachim.gossmann@gmail.com

Max Neupert
Bauhaus-Universität Weimar
Marienstraße 5
99423 Weimar, Germany
max.neupert@uni-weimar.de

ABSTRACT

We present an instrument for audio-visual performance that allows to recombine sounds from a collection of sampled media through concatenative synthesis. A three-dimensional distribution derived from feature-analysis becomes accessible through a theremin-inspired interface, allowing the player to shift from exploration and intuitive navigation toward embodied performance on a granular level.

In our example we illustrate this concept by using the audiovisual recording of an instrumental performance as a source. Our system provides an alternative interface to the musical instrument's *audiovisual corpus*: as the instrument's sound and behavior is accessed in ways that are not possible on the instrument itself, the resulting non-linear playback of the grains generates an instant remix in a cut-up aesthetic.

The presented instrument is a human-computer interface that employs the structural outcome of machine analysis accessing audiovisual corpora in the context of a musical performance.

Keywords

Audio-visual, Instrument, Real-Time, Concatenative Synthesis, Embodiment, Glitch, Cut-Up, Scatter Plot, Point Cloud, Video, Contact-free Control, *Leap Motion*

1. INTRODUCTION

Feature analysis can help to understand the characteristics and properties of a recorded media stream. It can pin-point events and highlight sections in the stream that become relevant to an observer. We can generate overviews, get an idea of the range and distribution of parameters and distinguish different elements, layers and chapters. We can classify and sort them, generate navigable catalogs to find them, et cetera.

With these contemporary tools for data analysis comes the question of their re-application to the activity of making music: How can we achieve intuitive access to the distribution of disclosed fragments, features and meta-tags for musical performance? How can we use the enormous amount of information the analysis has disclosed to generate new expressions? How can we provide a direct instrumental access to the re-assembly of media streams that enhances inter-human communication as opposed to confronting human beings with random or algorithmically generated pseudo-information?

Through collaborative research in technology and art practice, combined experiments and practical applications, we are investigating how we can use media fragmentation and feature analysis to *enable* the potential of human expression instead of *replacing* it by automatic re-assembly.

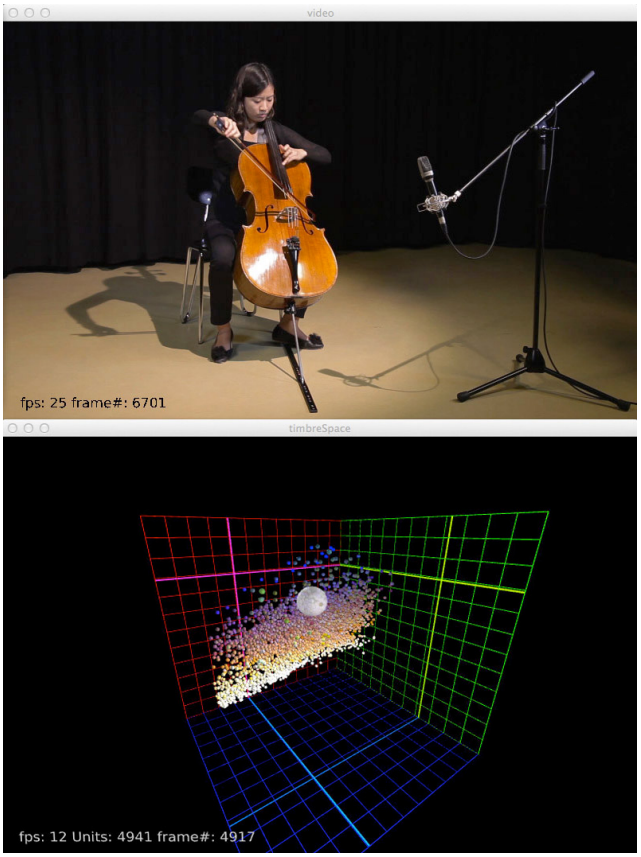

Figure 1: Top: video output in sync to the corresponding frames from the current playback unit. Bottom: pointer and units of the analysed recording scatter plotted in a 3D grid.

1.1 Instrument: platform for discovery and performance

In order to permit musicians to evoke specific perceptual experiences, musical instruments need to provide performers with a sense of accountability—reliable and predictable connections between player action and perceptual result.

How can we generate the potential for such accountability

NIME'14, June 30 – July 03, 2014, Goldsmiths, University of London, UK.

within the process of remixing fragmented media streams that are often complex and difficult to memorize? In his article *Transforming Mirrors* [11], David Rokeby describes how accountability in exploratory interactivity emerges in *belief systems* established by the perceived relationships between action and result.

Within this process of establishing accountability, *exploration* and *performance* fluidly merge into a single activity: Exploring the point cloud, the user gradually accumulates experience with the effects of their own physical actions on the perceptual result. A relationship between expectation and appearance unfolds that can sustain the interest of the player, while the disclosed range of possible action strategies gradually permits the player to *perform* effects intentionally. This process is not only an increase in the *control* over the behavior of the resulting playback, but following Alva Noë [9] we could say that *perceptual strategies* are established and gradually evolve during both the exploration of and performance with the instrument.

1.2 How the remix instrument connects to its users

The access a user has to the remix instrument has three perspectives. First, the strategy of fragmentation and analysis that is applied to the media content to generate the *corpus* has to be chosen and applied to the media stream. [1] In a second step, the analysis is visualized and transformed into an interface metaphor. The user can choose the specific mapping strategy by which the result of the analysis is transformed into a three-dimensional point cloud that can be interactively browsed to re-assemble the fragments. Finally, in a third perspective, the user is addressed with an audiovisual presentation: the playback of the re-assembled media stream. In the following paragraphs we will describe the current implementation of these components.

2. IMPLEMENTATION

2.1 Analysis and fragmentation

The instrument currently uses William Brent's *timbreID* [3] to decompose the audio component of a media stream into a collection of fragments (further also referred to as *unit* of the concatenative synthesis). Each unit is described by a feature vector in which each analyzed sound feature appears as added dimension in a multi-dimensional Cartesian space. The sound features analyzed are brightness, centroid, flatness, flux, irregularity, kurtosis, skewness, spread, zero-crossings, pitch, amplitude and BFCC1–BFCC7. A detailed description of the analysis strategies can be found in *Cepstral analysis tools for percussive timbre identification* [3]. The media stream in our example is a four minute audio-visual recording of a cellist. The musician has been instructed to create as many different sounds as possible by improvisation, thereby attempting to generate an encompassing *corpus* of the musical sounds that can be produced on the cello.

2.2 Visualization and navigation feedback

Every unit appears as an individual point. Together, all units of the recording visualize as a cloud. It forms a three-dimensional scatter plot that can be zoomed and navigated in real-time. Depending on which feature dimension is mapped on which axes of the cloud, the units appear in a different spatial distribution resulting in a different interactive accessibility (see bottom in figure 1). The varying visualizations of the analysis simultaneously help to understand the nature of the recorded sound and provide the user and player with an intuitive expectation of the unit's contents. Organization by pitch for example makes contained musical scales immediately visible. Correlations between different feature dimensions become apparent, for instance, when viewing the average pitch of a fragment versus the number of zero crossings its audio waveform contains. In that sense the remix instrument may also be used for Music Information Retrieval (MIR) as it permits the multidimensional analysis of a media artifact to be explored in a quasi-haptic way.

In order to trigger points inside the resulting point cloud we are using a non-haptic interface to define the right hand as the focal point of the playback. The virtual representation of the hand-pointer acts as a light emitter in the scatter plot, illuminating the distributed units and the Cartesian grid. Light, shadows and grid can help enhance the user's sense for the position of the pointer within the depth of the virtual space when it is displayed on a two-dimensional screen [7].

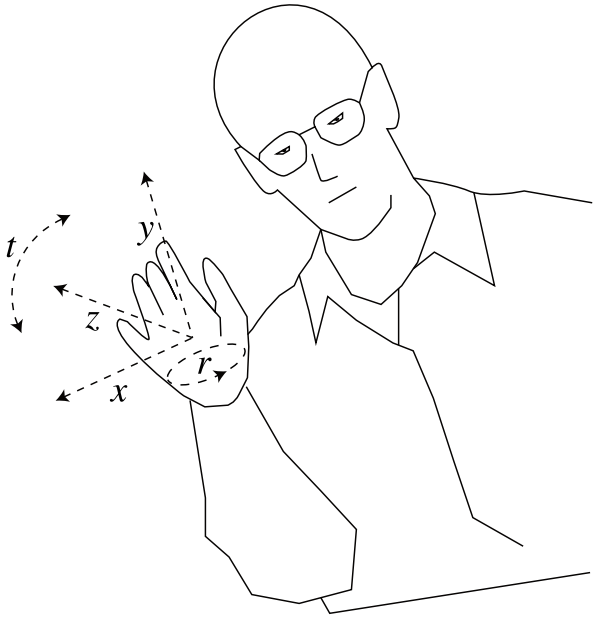

Figure 2: control parameters: position (x,y,z) roll (r) and tilt (t)

3. REMIXING

3.1 Human-Machine Interface

For navigation in the point-cloud we are using the *Leap Motion* controller. This consumer USB device is a compact infrared stereo camera with the necessary computer vision software running on the host. The software extracts position, tilt and roll for both hands and all fingers from the frames of the camera. Previously we had been using the *Microsoft Kinect* as HMI [8]. Both devices are contact free spacial sensing devices and have different capabilities. A brief summary of our findings comparing the two is as follows: The *Kinect* provides depth information for anything in the range of approximately 0.7-6m distance from the lens. It's output is a depth map which allows for a software based estimation of a skeletal structure also referred to as 'bones'.

[1] In the current setup we pre-analyze the media before playback to generate the analysis meta data that allows for indexing and display. On-the-fly analysis and indexing (of newly generated media) would be supported by the setup with minor changes to its structure.

The recognition of the skeletons fourteen joints' positions through the *Synapse* software is not instant and requires the user to stand in a frontal position, sometimes even a T-stance. When it recognizes a body once, the tracking of its joints is usually stable, but tracking of body and arm movements can be unreliable under certain circumstances, for instance when arms are close to the torso. The *Leap Motion* is designed to capture a much closer area and only tracks the hands. It generally has a lower latency and higher sampling rate than the *Kinect*. Access to the raw IR images is — as of today — not provided by the API. One can poll various data from the sensor, but only the following six values are of accuracy: palm position (x,y,z), tilt (x,y) and roll. Meanwhile data for the individual fingers is lost when they happen to be to close to each other or perpendicular towards the sensor. When a second hand is tracked the index of the hand (0,1) is not persistent to the right or left hand which makes this data unreliable to use. The latency doubles when a second hand comes into the field of vision of the sensor. With the developer tools comes a latency meter which can show a graph visualizing the latency on a rolling time axis. Since both devices use optical sensors they can't see that which is occluded from their vantage point. To conclude: both devices have their strenghts but also serious shortcomings. By choosing the *Leap Motion* over the *Kinect* we loose the ability to track the users head position but gain a higher precision and lower latency.

3.2 Interaction

The tracking of the user's hand allows one to access the point cloud. The performer's hand navigates the point-cloud and thereby selects particles to play back (see figure 2). Originally conceived as a two handed interface concept we now use only one hand. It's spacial position is directly mapped to the point cloud, while tilt is mapped to volume and roll to the grainsize. Like any instrument, it requires practice to accumulate experience with regard to the relationship between the user's self-movement and the resulting behavior of the instrument.

3.3 Performing concatenative synthesis

The re-assembly of sound from micro-fragments is usually addressed as *Granular Synthesis* [10] and more recently as *Concatenative Synthesis*, for example by Schwarz [13]. The re-assembly of media streams from larger units has also been addressed as *Musaicing* [1] and *Sound Spotting* [16].

While the waveform content of grains has been taken into account in certain re-assembly methods of *granular synthesis*, for example to enable a seamless combination of grains that avoids transition artifacts [2], the parameters used by granular synthesis to re-assemble a continuous sound from sound grains are generally agnostic to their specific content.

Concatenative Synthesis in the sense of Schwarz on the other hand juxtaposes a database of sound units — the *corpus* — to processes of algorithmic re-composition that operate on the descriptive metadata harvested in a process of signal analysis [14]. This allows the re-assembly of sound particles to relate to signal features present within the original media artifact. The reassembly of particles is organized along a *guide signal* for which appropriate units are selected from the database. This enables the creation of gestures within expressive dimensions that also characterize the various playing techniques of real instruments [15]— from noise-like to tonal, from low to high pitch, et cetera.

Our remix-instrument allows the performance of this guide signal coupled with the immediate perceptual result in the form of a continuous media stream assembled from the respective units stored in the database in real-time. By visually organizing the fragments or units along freely selectable analysis dimensions, the spatial layout of the *corpus* does not need to adhere to abstract parameters such as pitch transposition or particle length, but can be derived from signal features that are emergent *principal components* of the analyzed media artifact itself. Thus the performer of the sample-based remix instrument does not have to think in abstract synthesis parameters but is instead brought in direct contact with the structural aspects of the analytical *corpus* of available fragments.

3.4 Manipulated video

While the user navigates and performs the scatter plot and thereby links the sound units found in the displayed point-cloud, our system simultaneously shows the corresponding video frames. Analogous to the nature of concatenative synthesis in which non-consecutive units of the analyzed audio waveform are linked into a continuous stream, the video is played back alongside its respective audio fragment which results in random-access cueing, a *cut-up* of the original video footage: a real-time remix of audio-visual media fragments.

4. ADDITIONAL REMARKS

4.1 The corpus and 'instrument-inherent expressions'

As it may have become apparent to the reader, the purpose of our instrument is not necessarily to hide behind the content that was analyzed and fragmented in the creation of the corpus, but to instead use a performative approach to the corpus as a vehicle for new forms of expression.

We can explain what this means by a simple example: When regions of recorded sound are spliced together, the perceptual result necessarily lies on a continuum between two extreme cases:

1. the discontinuity can be imperceptible and undetectable. The experience of the listener is the illusion of a transparency toward the recorded material. This is the goal for example in the editing of classical music, where fragments from various recording takes are seamlessly assembled to create the impression of a continuous cello performance. The involved technology and assembly strategy become imperceptible.

2. the discontinuity becomes itself *expressive* to a degree that the original source of the sound becomes unrecognizable. The technological treatment of the fragment effectively severs the illusory connection between the observer and the sound source and the experience appears to emerge from *within the technological system.* In the realm of music, we find this strategy *to remove a sound from its source* exemplified in Pierre Schaeffer's Musique Concrete [12].

Perceptual artifacts generated within a technological system such as our remix instrument can alternatively be interpreted as disturbance or as an aspect of a specific expressiveness which the instrument makes available to the performer.

With strategies such as those presented in [2], artifacts of discontinuities between audio grains can be greatly reduced and may allow for the emergence of a *virtual cello.* The same is much more difficult to achieve for the video image. As a result, the corpus of the cello is currently rendered as a discontinuous playback of cut-up audio-visual fragments, an experience that balances between the material's origin in a cello performance and a distinct *instrument-immanent* experience.

Cello as source instrument appears especially appropriate for multiple reasons. On the one hand it offers an especially rich range of possible timbres and a wide frequency spectrum. On the other hand, it appears in the video image with roughly the same size as its player, which creates a balance in the perception of instrument vs. musician. Additionally, due to the seated position of the instrumentalist the movement is somewhat restricted and the instrument is usually at rest while the player moves around it in different ways, making the the cut up video easier to follow visually. Finally, we see the choice of the cello as an artistic reference to musician Charlotte Moorman who collaborated with artists such as John Cage and Nam-June Paik in experimental performances.

4.2 Perspective on particle clouds and embodiment

Different mappings between analysis parameters and interaction space evidently affect the intuitiveness of the interaction. Some mappings are more plausible than others, for example to map pitch to the y-axis, with the lowest at the bottom.

In (1.1) we have expressed a demand for reliable relationships between action and result that allow the user to develop from an explorer to a performer, and for the described interface to be *played* and performed as an audiovisual instrument—a process of embodiment as outlined by Paul Dourish [5]. We need reliable and stable connections between the interaction paradigm and the resultant audiovisual feedback, both within the visual interface itself and for the resulting playback of media fragments. This is especially important since our instrument does not provide haptic feedback like most traditional musical instruments, but instead relies on the comparably indirect visual feedback as well as auditory cues.

5. OUTLOOK

Future research may encompass the following aspects:

1. morphing strategies to fill the voids of the point cloud with interpolations between the closest data points,
2. inclusion of image data analysis similar to that described by Collins[4],
3. principle component analysis for the creation of point-cloud layouts of maximal relevance,
4. experimenting with other tracking infrastructures, aiming for higher accuracy and lower latency, like OpenStage or DTrack),
5. the use of a 3D screen or VR goggles to enhance the player's understanding of the scatter plot,
6. effects of different unit playback sizes and the application of unit-blending algorithms to allow for a greater control of perceptual artifacts occurring in the grain assembly,
7. collapsing higher dimensional clouds into the three-dimensional visualization using Multi-Dimenstional Scaling (MDS), as is implemented in the *Daphne Oram Browser* described in [6].

6. ACKNOWLEDGMENTS

Special thanks to cellist Hyunji Cho from *The Liszt School of Music Weimar*, Markus Westphal and Mario Weise of *Bauhaus-Universität Weimar* for the sound recording. This work is building upon *Pure Data* by Miller Puckette and its external libraries *Gem* by Mark Danks/IOhannes zmölnig, *timbreID* by William Brent and *shmem* by Cyrille Henry/ Nicolas Montgermont. It currently makes use of the leapmotion external by Chikashi Miyama built with the *flext* environment by Thomas Grill. All dependencies are open source.

7. REFERENCES

[1] F. P. Aymeric Zils. Musical mosaicing. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland*, December 2001.

[2] G. Behles, S. Starke, and A. Röbel. Quasi-synchronous and pitch-synchronous granular sound processing with stampede ii. *Computer Music Journal*, pages 44–51, 1998.

[3] W. Brent. Cepstral analysis tools for percussive timbre identification. In *Proceedings of the 3rd International Pure Data Convention, São Paulo*, 2009.

[4] N. Collins. Audiovisual concatenative synthesis. In *Proceedings of the International Computer Music Conference (ICMC), Copenhagen*, pages 389–392, 2007.

[5] P. Dourish. *Where the Action Is: The Foundations of Embodied Interaction*. The MIT Press, Oct. 2001.

[6] P. K. Mital and M. Grierson. Mining unlabeled electronic music databases through 3d interactive visualization of latent component relationships. In *NIME*, 2013.

[7] M. Neupert. Navigating audio-visual grainspace. In H. Reiterer and O. Deussen, editors, *Mensch Computer 2012 – Workshopband: interaktiv informiert – allgegenwärtig und allumfassend!?*, pages 319–322, München, 2012. Oldenbourg Verlag.

[8] M. Neupert and J. Gossmann. A remix instrument based on fragment feature-analysis. In *IEEE International Conference Multimedia and Expo Workshops (ICMEW), San Jose*, pages 1–5, 2013.

[9] A. Noe. *Action in Perception*. The MIT Press, Mar. 2006.

[10] C. Roads. *Microsound*. The MIT Press, Sept. 2004.

[11] D. Rokeby. *Critical Issues in Electronic Media*, chapter Transforming Mirrors: Subjectivity and Control in Interactive Media. Number 8. State Univ of New York Pr, June 1995.

[12] P. Schaeffer. Solfege de l'objet sonore, 1966.

[13] D. Schwarz. Concatenative synthesis: The early years. *Journal of New Music Research: Special Issue on Audio Mosaicing*, 3(35):3–22, 2006.

[14] D. Schwarz. Real-time corpus-based concatenative synthesis with catart – expanded version 1.1. In *9th Int. Conference on Digital Audio Effects (DAFx-06), Montreal*, 2006.

[15] D. Schwarz. The sound space as musical instrument: Playing corpus-based concatenative synthesis. In *New Interfaces for Musical Expression (NIME)*, 2012.

[16] C. Spevak and E. Favreau. Soundspotter-a prototype system for content-based audio retrieval. In *Proceedings of the 5th International Conference on Digital Audio Effects*, 2002.