Mapping to musical actions in the FILTER system

Doug Van Nort
Electronic Arts, Architectural Acoustics
Rensselaer Polytechnic Institute
Troy, NY
vannod2@rpi.edu

Jonas Braasch
Architectural Acoustics
Rensselaer Polytechnic Institute
Troy, NY
braasj@rpi.edu

Pauline Oliveros
Electronic Arts
Rensselaer Polytechnic Institute
Troy, NY
paulineo@rpi.edu

ABSTRACT

In this paper we discuss aspects of our work in developing performance systems that are geared towards human-machine co-performance with a particular emphasis on improvisation. We present one particular system, FILTER, which was created in the context of a larger project related to artificial intelligence and performance, and has been tested in the context of our electro-acoustic performance trio. We discuss how this timbrally rich and highly non-idiomatic musical context has challenged the design of the system, with particular emphasis on the mapping of machine listening parameters to higher-level behaviors of the system in such a way that spontaneity and creativity are encouraged while maintaining a sense of novel dialogue.

Keywords

Electroacoustic Improvisation, Machine Learning, Mapping, Sonic Gestures, Spatialization

1. INTRODUCTION

Our group, along with researchers in acoustics and cognitive science, have undertaken a project dubbed CAIRA, which stands for the creative, artificially-intuitive and reasoning agent [3]. This project is devoted to understanding and modeling machine performance from both a top-down, logic-based point of view well-suited to following rules as well as from a more bottom-up and intuitive approach to machine improvisation. created within this context, and stemming from an earlier pilot project with the same mission, is the the Freely Improvising, Learning and Transforming Evolutionary Recombination (FILTER) system [9]. This system places an emphasis on three key concepts: an embodied approach to machine listening, on intuitive and spontaneous transformations of a human performer's input in a way that is shaped by learning stylistic trends, and finally the system's actions are informed by an electroacoustic aesthetic that favors a sound-oriented view on performance output rather than one determined by music theoretic rules. Defining the space of possible musical actions for FILTER has been a fluid and integral part of the design process; the test-bed for this work is our trio Triple Point [11]. The details of the FILTER system related to machine listening and learning are discussed elsewhere [15], while here we focus on the design of the mapping to musical actions that results from continued use in improvisational sessions. A recent piece [14] is presented as an example of a musical context for which the system was adapted in response to particular performance demands.

NIME'12, May 21 – 23, 2012, University of Michigan, Ann Arbor.

2. MOTIVATION AND CONTEXT

The instrumentation for our improvisational trio Triple Point spans the spectrum of acoustic (soprano saxophone), acoustic modeling (Roland V-Accordion) and digital transformations based on analysis/resynthesis (GREIS system [15]). Through extended technique (saxophone), changing synthesis timbres (V-accordion) or by on-the-fly transformations (GREIS) our style is one in which sources can quickly fuse into a single element or conversely spread into unique, disjointed lines. Playing with source and instrumental identity in a manner that is dynamic and controllable is an important part of our music. Reflecting on the role of the GREIS player as one who listens for distinct lines of musical intention in the sound streams before capturing and transforming these, the design of FILTER was motivated to model this performance practice, adding to the richness of the musical interaction and allowing the possibility of expanding the group into a quartet. In addition to the GREIS system, the Expanded Instrument System (EIS) [15] is taken as a point of inspiration, with its focus on acting as a reactive, mirroring partner that re-presents past sound in a surprising manner in performance. Much like GREIS and EIS, FILTER has proven to be a unique partner that lends its own style and sonic character to our performance endeavors.

3. FILTER OVERVIEW

The FILTER system was designed to move beyond the notion of extending a performer's actions through time, towards a system that learns information that is embedded in the low-level structure of the audio stream of its improvising partner. While GREIS and EIS both have a running memory in the form of a recording of the past N seconds of performed audio on which to make decisions, FILTER encodes not only the waveform but also in parallel the fine structure level of information about the temporal evolution of sound features. In parallel with this low-level information the system catalogs a set of *sonic gestures* which give semantic meaning to performance actions at the note and phrase level. The details of listening and learning are discussed elsewhere [15], though a brief overview is required in order to understand the parameters that result from this stage as they are mapped to and directly determine the output musical actions.

3.1 Listening, Learning

The structuring principles for this aspect of the system are that of listening to *gestures* and *textures*. The former can be

thought of as foreground actions that have a coherent motion in regards to spectrotemporal parameters, while the latter is the characterization of the overall sound field over a larger time duration (e.g. larger than 5 seconds), in the absence of coherent motion. Listening for and recognizing gestures is based on continuous gesture following [2] applied to a small set of sound features. In contrast to many applications of gesture following applied to sound streams, in FILTER this is based on unsupervised learning as follows: when a transient is detected in amplitude or fundamental frequency, subject to an inter-onset temporal threshold, a new "sonic gesture" is considered to have begun. If this gesture is dissimilar to anything in the current "gesture space" then this may be added as a new member of the space, with an older one possibly being discarded. This aspect of learning can be thought of as the system developing a semantic memory and deciding which are the relevant sonic gestures on the fly, in the non-idiomatic spirt of free improvisation. The output of this process is a continuous likelihood that a given action is related to one of the sonic gestures in the given space, thereby providing a continuous degree of certainty that the system is hearing those sonic gestures that have been internalized within the performance moment.

This learning stage is out-of-time in the sense that each gesture is committed to FILTER's semantic memory without any temporal ordering between gestures. In parallel with this, the underlying audio's temporal structure is learned in order to provide an understanding of the temporal regularity and similarity across an entire performance. This episodic memory component of the system is inspired by the audio oracle concept [5] that underlies the OMax system [1], and FILTER utilizes this project's Max/MSP implementation of the factor oracle algorithm. From experience we recognize and leverage the power of this learning algorithm, but also see the shortcoming of requiring a human operator to make contextual choices when using this information for musical playback. We construct the semantic-level gesture/texture listening layer and use this to drive output actions to explicitly address this issue. The episodic and semantic levels of information in FILTER are associated by endowing relevant states with time-stamps for any associated sonic gestures in memory, as well as segment boundaries between areas that are considered to have different textural qualities [13].

3.2 Evolving, Mapping

The use of analysis, recognition and structure-learning in FILTER is not towards the end of categorically classifying performer actions in an out-of-time fashion, but rather the goal is to leverage the process of recognition and understanding as it develops, and so the system focuses more on *anticipation* than on recognition. As between human improvisers it is the feedback and reinforcement of certain sounds or passages, guided by moment-to-moment anticipation, that give them meaning within a given performance context. Further, the design of FILTER is predicated on the idea that this continuous recognition should not lead to output that always exists within the same parameter space, but that this should shift over time in a way that favors novelty and challenges a human performer, though not so erratically that it is perceived as random - a balance of spontaneity and the desire for dialogue. With this in mind, the system output is partially governed by an evolutionary process that acts as mediating layer between the learning elements and the space of output parameters. Building on a previous project [12], FILTER maintains a notion of continuous, dynamic attention that is tied to a measure of *confidence* in the gesture listening, defined as:

$$C_n = \delta(m_n - m_{n-1}) \sum_{k=0}^{n} 2^{\frac{-1}{\lambda_k}} (m_k d_k)$$

where m_k represents the value of the maximal likelihood for the kth gesture in the gesture space and d_k is the deviation from the average value, also for the kth gesture. The binary function δ is present so that if there is a sudden change in gestural probabilities, the confidence value is zeroed before again rising. This allows the system to follow stable gestures, but also to adapt to a perceived sudden change in musical direction. The smoothness of the confidence measure is tunable by the values λ_k. In order to allow FILTER to move towards a globally predictable direction while maintaining random elements on a local scale, a genetic algorithm (GA) is used as a layer between gestural listening and the space of possible behaviors. What sets this usage apart from many projects related to evolutionary music [6] is that this is not an interactive GA implementation wherein the user explicitly rates the goodness of each output – which is a substantial time and attention bottleneck. Rather, the fitness is directly tied to the saliency of the gestural recognition process by mapping the smoothed gesture-likelihood values into the fitness of a member of the GA pool, while the confidence is inversely proportional to the mutation rate. In this way, the "goal" changes as a product of the system's gestural recognition and confidence. However, if the confidence remains substantially low then the *mode of listening* for FILTER changes so that the gesture-based confidence value no longer drives the GA, and instead the texture-based sound features influence the output to system behaviors.

There are two layers of mapping in this part of the system. The first is the association of members of the gesture space, as well as textural categories, into output behaviors. This is achieved by mapping archetypal gestures - defined by their temporal shape or morphology [7] - into each member of the GA population on one hand, or an archetypal set of texture features (averaged over a large time window) on the other. This mapping provides a semantic association to begin (e.g. "repeated sharp attacks should give rise to X type of behavior"), which can be thought of as a set of musical values for the system. These 'value mappings" are reinforced or lost over time as the parameter space evolves and FILTER is influenced by the style of its improvising partner, though they may be reinstated during the course of performance. The second layer of mapping is the embedding of these GA members in a continuous, higher-dimensional space of possible behaviors through the use of continuous mapping strategies [10]. In this implementation, a set of N-dimensional population members move within a simplicial complex where each node is associated with a behavior state of the system. The member of the population associated with the currently most salient gesture/texture state is used to interpolate the nodes of the enclosing simplex, determining an output behavior state of the system. This can be thought of as a cloud of possible states that move with a quasi-physical nature as determined by the output of the listening module. In this particular aspect, FILTER shares a similarity with the continuous state-based approach of the Ozone project [8]. One critical difference is that the current state of the system behavior itself (the so-called "stability" feature) also partially determines the movement in the state-space – a sort of self-reflexivity of the system.

4. MUSICAL BEHAVIORS

In FILTER, the learned graph-like audio structure is navigated in order to produce sound output. Navigating this

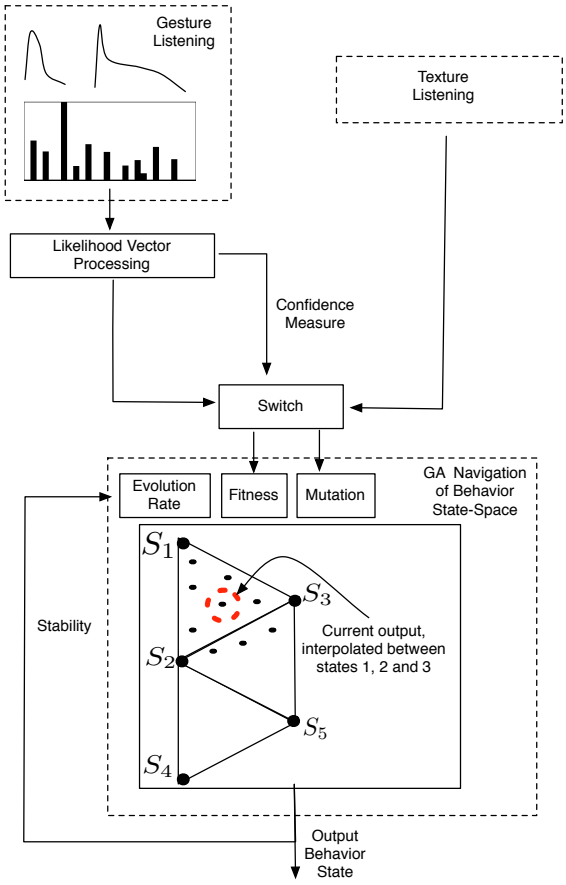

Figure 1: FILTER's Listening system balances between gesture/texture listening, with the output causing an evolution of members within the interpolated behavior space.

structure causes the system to recombine past elements of audio that have varying degree of contextual relevance. By reinterpreting the nature of this "relevance" on the fly and altering the manner of recombination, FILTER moves beyond the aforementioned problem of requiring a human operator. Coupling this with additional sound transformations gives FILTER a set of potential performance re-actions that are conducive to Triple Point's sound-oriented, free improvisation aesthetic.

The behavior states of the system give a high-level description that is then mapped into musical actions, as well as into internal decision making. As noted in figure 1, this mapping is partially regulated by the relative saliency of gestures (vs. textures) to the current musical context. The continuous behaviors include:

- Rhythmic-ness: The likelihood that individual lines will be repeated, as well as the degree of variation within a given repetition.
- Wildness: If gesture listening is dominant, the likelihood that the system will mirror the performer by using recent input vs. improvising on disparate regions of past and present input. If texture listening is dominant, the likelihood that the system will draw on past regions that have similar textural sound qualities.
- Stability: The likelihood of possible change in the overall behavior state of the system.
- Sustain: The favoring of sustained vs. short tones or actions.
- Density: If gesture listening is dominant, this affects the size and spacing of output phrases. If texture listening is dominant, the number of overlapping layers of content that are performed at once.

Note that the function of these behaviors changes depending on the listening context. Further, the wildness state determines the likelihood that the system will behave similarly or differently from the player, whether this is "gesturally" (a single, well-defined passage) or "texturally" (layers of quasi-repeated or stretched passages of sound). Therefore the notion of same/different playing has the dual interpretation of drawing on similar/different content, or playing in a similar/different style.

4.1 Transformations

In addition to recombining disparate fragments of audio in a manner subject to the given behavior state, FILTER has the ability to time stretch and pitch shift its current musical output, or each layer individually in the case of denser textural playing. Each of these potential phrases may also be fed into a feedback delay line that is subject to modulation and filtering. These fundamental processes define a variety of musical effects, as determined by the mapping from the higher-level behavior parameters and by which listening context is currently dominant. The degree of similarity in playback style - expressed by the Wildness state - also determines the amount (if any) of pitch shifting, while the degree of sustain influences the amount (if any) of time-stretching applied. The set of transformations defined by feedback and modulation are influenced by both wildness and stability in a cross-coupled fashion.

4.2 Spatialization

An integral part of the system's musical actions is its ability to define spatial gestures that react to the musical context. The system utilizes the virtual microphone control (ViMiC) approach [4], which models sound reflections, dispersion patterns of sound sources and doppler shifts. As such it is very conducive to rapidly moving sound sources around the space in a realistic fashion, where relative positioning between source and speaker output may be controlled. The parameters that are subject to machine control in FILTER include: the set of possible trajectories for each sound source, the reverberation decay time, room size of the spatial model, and the radius, speed and incidence angle of each sound source. These spatial parameters are given equal importance to all other musical actions/transformations in consideration of the overall performance of the system. For example, the textural nature of the output is drastically altered if each improvised line of the system is presented as a different moving source, thereby separating each one spatially. This interaction between spatial gesture and machine actions was the subject of consideration in a recent telematic piece that we presented at last year's NIME conference, which we now describe as an example application in a real musical context.

5. DISTRIBUTED COMPOSITION #1

The presentation of a work involving FILTER and Triple Point is a challenging (yet rewarding) endeavor as it involves multiple complex systems. In Triple Point there is already a sharing of sonic gestures through the capturing and transformation of audio on-the-fly (Van Nort, GREIS) that extend the overall sound scene. Presenting the actions of FILTER so that they exist as a unique contributor to the musical dialogue presents an interesting challenge. In the piece Distributed Composition #1 we embraced this

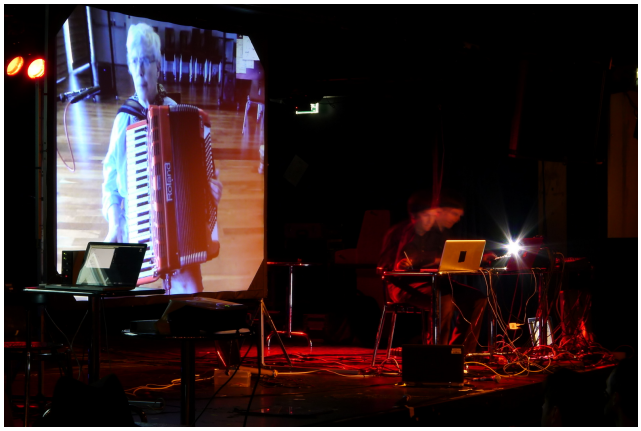

Figure 2: Performance of Distributed Composition #1 at NIME 2011 (Photo by Alexander Refsum-Jensenius).

complexity and pushed it further by defining a three-site telematic piece. The title of the piece refers not only to this physical distribution of the human players, but also to the distributed musical cognition between human and machine, as well as the fact that each player had a hand in defining the musical structure – making it a distributed composition in several senses of the word. The FILTER system itself had a hand in composing the structure in that it acted as conductor, determining when a member of the quartet would have the option of playing. This was achieved by adding these cues to the behavior state-space, while an audio matrix determined which input FILTER was improvising on at a given moment. Within these confines, any of the eligible four players were free to improvise. The piece allowed the FILTER system to capture the GREIS output as well (while both were capturing the remote acoustic players), resulting in a proliferation of certain phrases that were subject to several iterations of musical transformation. The staging and sonic display for all human and machine players was adapted so as to allow for a more coherent musical dialogue in light of this sharing of sources. First, the local and remote human players were presented on stage (see figure 2), and their sound was localized to the stage. At the same time FILTER was only present in the surrounding eight channels of audio. Secondly, the system was populated with a set of musical values such that the particular palette of gestures used by the GREIS player for the piece would lead, with high likelihood, to actions that were considered quite different from the current musical context set by that player. In practice, this often led to the FILTER system performing in a very stable, sustained and spatially distant fashion when the GREIS player was producing sounds that were full of transients. Meanwhile when the system improvised on the content from the acoustic players, the result was often a rapid spatial gesture that moved with a small radius in the center of the space, a "musical value" that was added so as to help subvert the distance one might feel in a telematic presentation such as this.

From the experience of this piece, we feel that intelligent, reactive spatial gestures that are integrated into the musical context are a very fertile area of exploration in the case of telematic performance in particular. Further, our hope is that this piece can serve as an early, novel example in terms of a distributed approach to composition - as a non-hierarchical mode of engagement and planning between human performers as well as between human and machine performers, where each is potentially located in disparate regions of the planet.

6. CONCLUSIONS AND EXTENSIONS

FILTER has proven to be a convincing improvising partner, not only in use with Triple Point but in performance with a number of players on bassoon, piano, cello, violin, electronics, various percussion and other instruments. Allowing the system to be flexible in the sense of redefining the high-level mapping to values and behaviors is key to the system's musicality, as with any human performer wherein one can discuss musical intentions a priori. At the same time, the fact that the design allows for a considered coupling between analysis, recognition and evolved output parameters is an important part of why the system remains convincing and reliable. While the system is under continued development with its current mission and electroacoustic aesthetic in mind, work is currently under way in a parallel project with colleagues in cognitive science (under the CAIRA intiative) in order to explore the result in the case where FILTER is submitted to decision making that is a product of a logic-based reasoning module which acts on long-term information, related to musical tension.

7. ACKNOWLEDGMENTS

This project received support from the National Science Foundation (#0757454 & #1002851).

8. REFERENCES

[1] G. Assayag, G. Bloch, M. Chemillier, A. Cont, and S. Dubnov. Omax brothers: A dynamic topology of agents for improvisation learning. In *ACM Multimedia Workshop on Audio and Music Computing for Multimedia*, 2006.

[2] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. *Lecture Notes in Computer Science (LNCS), Gesture in Embodied Communication and Human-Computer Interaction*, chapter Continuous realtime gesture following and recognition, pages 73–84. Springer Verlag, 2010.

[3] J. Braasch, , D. Van Nort , S. Bringsjord, P. Oliveros, A. Parks, and C. Kuebler. CAIRA - a Creative Artificially-Intuitive and Reasoning Agent as conductor of telematic music improvisations. In *131st Audio Engineering Society Convention*, 2011.

[4] J. Braasch, N. Peters, and D. L. Valente. A loudspeaker-based projection technique for spatial music applications using virtual microphone control. *Computer Music Journal*, 32(3):55–71, 2008.

[5] S. Dubnov, G. Assayag, and A. Cont. Audio oracle: A new algorithm for fast learning of audio structures. In *Proceedings of the 2007 International Computer Music Conference*, 2007.

[6] E. R. Miranda and J. A. Biles, editors. *Evolutionary Computer Music*. Springer, 2007.

[7] G. Peeters and E. Deruty. Sound indexing using morphological description. *IEEE - Transactions on Audio, Speech and Language Processing*, 3(18):675–687, 2010.

[8] X. W. Sha, M. Fortin, N. Navab, and T. Sutton. Ozone: continuous state-based media choreography system for live performance. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1383–1392, New York, NY, USA, 2010. ACM.

[9] D. Van Nort. FILTER. `http://www.dvntsea.com/projects.html`. *Last Accessed April 20, 2012.*

[10] D. Van Nort. *Modular and Adaptive Control of Sound Processing*. PhD thesis, McGill University, 2009.

[11] D. Van Nort. Multidimensional scratching, sound shaping and Triple Point. *Leonardo Music Journal*, 20, 2010.

[12] D. Van Nort, J. Braasch, and P. Oliveros. A System for Musical Improvisation Combining Sonic Gesture Recognition and Genetic Algorithms. In *Proceedings of the 6th Sound and Music Computing Conference*, pages 131–136, Porto, Portugal, 2009.

[13] D. Van Nort, J. Braasch, and P. Oliveros. Sound texture analysis based on a dynamical systems model and empirical mode decomposition. In *129th Audio Engineering Society Convention*, 2010.

[14] D. Van Nort, P. Oliveros, and J. Braasch. Distributed Composition #1. `http://www.nime2011.org/program/concerts/`. *Last Accessed April 20, 2012.*

[15] D. Van Nort, P. Oliveros, and J. Braasch. Developing Systems for Improvisation based on Listening. In *Proceedings of the International Computer Music Conference*, New York, 2010.