DEVELOPING SYSTEMS FOR IMPROVISATION BASED ON LISTENING

Doug Van Nort, Pauline Oliveros, Jonas Braasch
School of Architecture and Department of the Arts
Rensselaer Polytechnic Institute, Troy, NY
vannod2@rpi.edu

ABSTRACT

In this paper we discuss our approach to designing improvising music systems whose intelligence is centered around careful listening, particularly to qualities related to timbre and texture. Our interest lies in systems that can make contextual decisions based on the overall character of the entire sound field, as well as the specific shape and contour created by each player. We describe the paradigm of “expanded instrument” systems that we in turn build upon, endowing these with the ability to listen, recognize and transform performer’s sound in a contextually relevant fashion. This musical context is defined by our free improvisation trio, which we briefly describe. The different modules of our current system are described, including a new tool for real-time sound analysis and method for sonic texture recognition.

1. INTRODUCTION

When a group of musicians engages in free improvisation, there is no a priori set of rules that the players use to communicate their intent. Rather, the musical language that players develop is acquired throughout the course of a performance or rehearsal session. While particular harmonic, melodic and rhythmic structures might be presented from various styles, the intent behind this presentation is completely open and unknowable beforehand, meaning that such intent must be found through careful listening to these qualities but also more fundamentally to the nature of the sound itself. Therefore, understanding the intention behind a sound event arises from careful attention to the shape and trajectory of salient sound features, the state of the overall sound field and a memory of recent actions taken by fellow improvisers. This need to define “style” as a function of the immediate musical situation has prompted Derek Bailey to refer to free improvisation as “non-idiomatic” [2]. It has been our experience as improvising musicians that structuring the musical interaction by fundamental listening principles – rather than musical stylistic rules – is an effective way to engage in the act of free improvisation. We therefore have focused our work in interactive system development on building an involved machine listening system that is coupled with musical decision making, and is designed with certain modes of listening and musical reaction in mind.

In particular our work is strongly informed by the practice of Deep Listening [11], including the notions of focal and global modes of attention. The former is concerned with discovering a particular sonic event and following this as it shifts over time. The latter is concerned with listening to the entirety of the sound field – everything that one can perceive. The key to our use of these principles is that it is not one or the other, but rather the process of these modes, and the dynamics of balancing the two that are embedded in our work. Meanwhile, a second structuring principle arises from particulars of our musical aesthetic. Our interest is in a blending of acoustic and electronic sound such that source and cause of an event can be obscured, where gestures are traded between musicians within the sound field. Achieving a structured listening in such performance practice – with a necessary level of abstraction and receptiveness – provides similar listening challenges to electroacoustic music, and we find e.g. Denis Smalley’s principles of Gesture and Texture [17] to be appropriate metaphors for further defining our system’s listening process. The former refers to an “energy-motion trajectory” that provides a sense of human intention and direction, while the latter refers both to the inner detail of a given sound event as well as the quality of the entire sound field. Our work is based on the conviction that these principles of focal/global and gesture/texture provide an excellent framework for developing a system based on a close coupling of listening, recognition and creative interaction that speaks to the self-organizing and emergent nature of form in free improvisation.

2. RELATED WORK

In the realm of real-time interactive systems, there have been many examples of work that is focused on creating an autonomous partner for musical improvisation. One very prominent system is George Lewis’ Voyager [9], which converts pitch to MIDI data in order for various internal processes to make decisions depending on the harmonic, melodic and rhythmic content before producing symbolic output to interact with the human performer. A more recent system that focuses on timbral parameters as well as analysis of
their gestural contour was taken by Hsu [8] in which parameter curves for pitch, loudness, noisiness, roughness, etc. were extracted, with captured sequences being stored in a database. Our system shares the conviction that *sonic gestures* [20] are essential to recognize in human-machine interaction for improvisation. However we differ in that the recognition itself is on-line with our system, and continual adaptation to the *anticipated* gesture is used as an element of the machine intelligence.

Taking off in a different direction, Pachet [14] has focused on creating a system that is both interactive and which learns musical style. This approach is based on variable Markov chains that learn a tree structure through analysis of input MIDI data. The goal is focused on imitation and predicting musical sequences in a given style. The author claims that the system can “fool” a listener on the short term, but fails to convince on the scale of an entire piece. A similar approach is found in the OMAX system [1], which learns variable-length musical sequences, and can understand phrases as being recombinations or variations of past stylistic tendencies. These approaches are interesting in their lack of style assumptions a priori. However, the ability to produce “creative” output is more crucial to our work than a system that presents variations on a current theme, and so we approach such methods as one part of the learning paradigm to be augmented by creative and contextual decision-making. On the other edge of the “novel output” spectrum, Blackwell and Young [5] have focused on the use of evolutionary techniques such as a swarm algorithm to guide the spontaneity of the improvisation. Recognizing the benefits of these two approaches, we utilize a combination of similar techniques, structured both by the aforementioned listening as well as our particular approach to engaging with the computer through transformation of the performer’s audio in real-time.

3. PARADIGM OF EXPANDED INSTRUMENTS

The Expanded Instrument System (EIS) [13] began at the end of the 50s, and was informed by listening to the latency between the record and playback heads on reel-to-reel tape recorders. This monitoring of signals led to improvisatory manipulation of amplitudes, which in turn afforded changes in sound quality. Using feedback within and between tracks on a stereo machine increased the possibilities, while adding a second machine with a shared tape afforded a second and longer time delay, and more routing possibilities as described in [11]. Though limited by the available hardware, through control of time delays, amplitude and signal routing this rudimentary EIS produced music with changing timbres and spatial qualities such as Bye Bye Butterfly and I of IV [12], all improvised in the studio by the composer.

Improvisation, listening and system development continually informed and increased the sophistication of EIS. In 1967 musical instruments were included as sound sources, allowing them to be colored by delays, layering and volume changes. Moving away from unwieldy tape machines, in 1983 tape delay was replaced by digital delay. Modulation of the delayed signals was introduced so that a signal could be processed and returned with a new “shape”. Delayed signals could also now be shaped via voltage control by a foot pedal. These innovations led to the desire for more delays and the need for a better performance interface in order to control the many signal paths [10]. With the transition of EIS to the computer [7], continuous foot gestures remained an integral control paradigm while a variety of machine-driven modulations were programmed in order to shape the sound in tandem with this performer action. These same modulatory gestures were similarly applied to spatialization of the sound output, as EIS was also conceived as a multi-channel system and gestural movement of sounds is an integral part of the system aesthetic.

In 2001, algorithmic control of delay times, modulations and spatialization was implemented. This allowed for meta-control of up to 40 delay times having fluctuations from milliseconds to one minute, as well as type and depth of modulations. This higher-level of control allowed a performer to create a sparse to dense field of moving sound that demanded focal and global listening, as each new sound would exist in some form in the future and also become a part of the past. In this sense, EIS is a “time machine” allowing the performer to expand, listen and react – making improvisatory decisions to all three times and to multiple parameters simultaneously. The performer reacts to material that is a combination of human and machine modulations. EIS has always acted as such an improvising partner, pushing the performer to make instantaneous decisions and to deal with the consequences musically.

3.1. EIS: Present and Future

Currently EIS exists with algorithmic controls that can be set to learn and repeat patterns of control data, modulation forms can be drawn and edited by the user, as can spatialization patterns. Nevertheless, performer control on the fly is still daunting, and higher-level control is continually being re-considered. Changes in EIS parameters are governed by 50/50 randomization, a chaos generator or a random event generator – and so the performer is challenged to develop ever more extensive abilities of intuition regarding what to play and when to play it. The result is a fluctuating mix of live instrumental input with machine feedback that has the feel of human partnering, mirroring his/her own sounds interactively.

The current delay processor module has up to twenty delays. Modulator types may be selected by the user, turned off or changed randomly with user control for the rate of switching. The number of active delays may be set manu-
A second and more profound difference is that GREIS has focused on the notion of “scrubbing” captured sound with the dominant hand as it enters one of the system buffers, while modulating the output with pen articulations and left hand control. This approach extends the notion of the computer-mediated modulating delay lines of EIS, which can be thought of as a turntablistic “scratching” of the captured audio, with a sound that ranges from a gentle phasing to a dramatic screeching, both of which alter the pitch content as they provide a new gesture in response. The use of scrubbing in GREIS returns the modulation to the musician (whose hands are free in this context), so that material can be shaped into a new sonic gesture through capturing sound and re-shaping amplitude envelope and other features over time. This is furthered by the use of a combined granular and time/frequency approach that allows for extreme time stretching and pitch shifting in a decoupled fashion, but also allows for control of the textural nature of the signal in a dynamic fashion, placing focus on a combined gesture/texture shaping.

The system achieves this control through treatment of “grains” in a general sense to mean both temporal or spectral units that many be shaped by the tablet and propagated through the system in a large feedback delay network. This is achieved through a combination of two granular synthesis modules and two phase vocoder based modules. The granular technique introduces two separate processing channels: one being controlled by an MSP phasor~ object at signal rate, and one being controlled by a Max metro/gate system at control rate. This allows for a very smooth stretching with minimal artifacts in the former case, while the latter allows for grains to be separated intro streams with each being processed in a unique way. In GREIS, each of these temporal grains has a feedback delay and pitch shifting applied (allowing for a very profound texturization) before each individual stream is sent into the primary feedback matrix (lower right of figure 2). The control-rate grain streams are configurable to 8, 16 or 32 separate streams per processing module. The signal and control rate grain channels can be mixed separately, and an involved mapping structure is used in order to transition between processing types and synthesis parameters, resulting in the ability to either hide or accentuate digital artifacts [19]. The “spectral grain” processing is based on a combination of gabor modules and the super phase vocoder [16]. This allows for partials, noise and transients to be separated in real time as the sound is scrubbed, and each of these are propagated through the primary system matrix as well. Important to GREIS’ ability to shift between different sonic shapings and textural transformations is that the control interface remains the same while only the state of “temporal” or “spectral” is changed (in the three identical modules of figure 2, this is reflected in the upper left drop-down menu).

The final key aspect of the system is that it merges this “hands-on” approach to shaping new sonic gestures (while
in the process breaking apart the textural components) with the EIS approach to computer-mediated gestural modulations. The large-scale modulations are thus placed in the feedback matrix, allowing for an interplay between human-based shapings that leverage the power of time stretching and parametric resynthesis with surprising new delay-line modulations that further reshape the already-transformed sound, arriving at unsuspecting times. This delayed output can then pass back into the granular-feedback chain to be further reshaped and remain in memory longer as a new sonic object. Either multi-dimensional Wacom tablet gestures or the resultant scrubbed audio may be saved in memory, and the performer can recall these at a later point in performance. Finally, a higher-level control layer of directed randomness introduces noise into the sound processing parameters as well as the matrix routing and per-grain filtering, while the behavior of the modulated delay-lines may be controlled probabilistically.

3.3. Towards an Intelligent Response

In both EIS and GREIS, many surprising sounds can arise that are a re-presentation of performer actions. These results are most often an interesting commentary on the sound quality as well as the musical structure. Both systems (and particularly EIS) have been used in countless recordings and performance contexts, and are regarded as having high degree of musicality and interesting musical interaction potential by those who have performed with the systems. Having said this, both rely on a performance memory and “decision making” that arise from a well-tuned (in regards to range, timing) set of random generators, whose effectiveness lies in having a behavior that can be anticipated and manipulated by performers. Therefore while the system presents the musical material in a new context to be considered by performers, the system does not listen to the sonic and musical context, to present material in a new way as a result of this. Such intelligent decision making, based on our listening principles, is what we strive to introduce. Some of our designs to this end will be discussed from this point forward, after describing the performance context that provides the musical “problem” and source of “data” that inspires and directs our system improvements.

4. TRIPLE POINT

Triple Point (figure 3) is our improvising trio whose core instrumentation is GREIS, soprano saxophone and digital accordion synthesizer. The name refers to the point of equilibrium on a phase plot, which is a metaphor for how we create dialogue as performers. Our musical interaction is centered around an interplay with proper acoustics, modeled acoustics (from the Roland V-accordion) and electronics. The GREIS performer captures the sound of the other players in real time, either transforming these to create new sonic gesture or holding them for return in the near future. The V-accordion player changes between timbres and “bends” the presets of the instrument through idiosyncratic use of the virtual instrument, while the saxophonist explores extended technique including long circular-breathing tones and multophonics. This mode of interaction has resulted in situations where acoustic/electronics source is indistinguishable without careful listening, while other times this becomes wildly apparent. This morphing is based on timbral transformations of the “acoustic” players, but just as much is a product of these players playing into the process of transformation as it occurs. Examples of this performance style can be found
improvisation by attempting to recognize intent. We search
for the most likely type of sonic gesture that is being played,
with a continuous degree of certainty about this understand-
ing. We are less interested in a musical retrieval than using
the process of musical retrieval in a way that an improvising
performer does, continually updating their expectation.

5.1. Feature Extraction

Our choice of features is focused on the ability to differen-
tiate between the sound made by a performer along pitch,
dynamics, timbre and texture dimensions. We further need
features that work well for the wide palette of physical mod-
els from the V-accordion, but also that are sensitive to the
rich acoustics from extended saxophone technique. At the
same time these features need to make sense as a sequence
of states for the resultant HMM model. We have found that
global spectral features related to timbre do well for a gross
characterization of different gestures, while finer qualita-
tive differences can be distinguished by looking at features
that react to textural differences, which can be considered
as spectro-temporal features that are more separable in time
and acting over a larger time scale than timbre (e.g. less than
20ms vs. 20-400ms). Rather than pitch we look more gen-
erally at “pitchness” and temporal regularity. These classes
of sound features are thus decomposable as global spectral,
pitch strength and textural. To this end we have found spec-
tral centroid and spectral deviation most effective (in our
trio) for the first group, frequency, energy, periodicity and
ratio of correlation coefficients derived from a normalized
autocorrelation pitch model for the second, and finally we
use the 3rd and 4th order moments of the LPC residual for
the final category. These react well to the “inner” textu-
ral detail, while the approach in section 6 detects the global
texture of the overall auditory scene.

5.2. Sonic Gestural Analysis

The initial process of recognition is triggered by onset and
offset messages, using an adaptive threshold version of the
onset detection method described in [3]. This operates in
the complex spectral domain to provide a continuous onset
function that can be thresholded for on/offset. The system
can look at sharp changes in any of the above features, and
we have found that detecting onsets in dynamics, fundamen-
tal frequency and spectral centroid are the most effective
measures for possible event segmentation. The detection of
such an event triggers the recognition process, but the final
determination of an event’s significance is the result of the
recognition stage. This component is built using combined
HMM/dynamic time warping [4] and the FTM package for
Max/ MSP.

This recognition system is trained on example sonic ges-
tures, which are represented as a left-to-right topology of
states, each comprised of the multidimensional feature set.

Figure 3. Triple Point Performing at Roulette in New York.

on the album Sound Shadows [18].

The blending of these disparate sources tends to have
two distinct modes: when separate gestural contours arise
that are distinct with sources that are difficult to follow, and
a fusion of sources into a singular texture that moves in a co-
herent fashion, or which lacks a linear direction. This, again,
is guided purely by focal and global listening in the moment,
and is both an awareness of the motion between gestural dy-
namics and textural sustain. The group decides to follow
these recognizable musical states of: striated gestures, ges-
ture/texture mixture, complete texture or disjointed sources
(a sort of “noise” in this context). This observation on our
playing style has motivated several sub-systems for machine
listening and analysis that we have since endowed with addi-
tional sound transformation capabilities, resulting in an “in-
telligent agent”. While Triple Point at its core is the above
trio instrumentation, we at times also become a quartet or
even quintet by including EIS or the agent as additional per-
formers. These extra performers are in turn noted on pieces
or recordings, in order to allow the listener to navigate the
interaction and intent behind the spectra of human/machine
and acoustic/electronic at varying levels of sonic and musi-
cal structure.

5. RECOGNITION OF SONIC GESTURES

The above mode of interaction and set of musical states are
a product of inter-performer interactions but also of the ges-
tural modulations presented by GREIS +/- EIS, and so to
invert this process we look to build a system that recognizes
different sonic gestures. This aspect of the system, first pre-
sented in [21], begins with analysis of sound features fo-
cused on combined timbre and texture recognition, followed
by a Hidden Markov-Model (HMM) based approach. While
HMMs have become commonplace for speech recognition
and even recognition of musical timbre, most often the inter-
est lies in an out-of-time or a posteriori act of classification.
By contrast, our work specializes to the uncertainty of free
This defines a sort of gestural “dictionary” that all incoming gestures can be compared against. These gestures should be orthogonal in some musical sense, which is taken care of either a priori by exemplary training, or by on-line learning as described in section 7. Once the system is trained on a set of gestures, it compares any segmented input gesture to its internal lexicon of gestural archetypes, providing a vector of probabilities for each member. While a standard classification system would look to the most likely member, we in fact examine the entire probability vector and its dynamic form as it changes continuously in real-time. This information is key to defining dynamic attention, as it is used to represent the system’s expectation of what a sonic gesture is at any given moment, which in turn is continually being updated.

In particular, we extract the normalized probability, the maximally-likely gesture and the deviation between the maximum and the few highest values. These latter values each give some indication of how strongly the system believes that the performed gesture is one from the system. Both values are needed in order to know uniqueness as well as strength of recognition. In addition to these instantaneous recognition values, we define a confidence measure based on a set of leaky integrators applied to the maximum and deviation values [21], which reflects a building up of confidence in the likelihood of a given sonic gestures over time. Therefore, the system can change its mind abruptly, or become more certain over time in a continuous fashion. Changes in system output follow accordingly by mapping these extracted probability features into sound transformations or higher-level musical state changes.

The final key component of this sub-section of our listening agent is that we utilize several gestural spaces in parallel, that each cover distinct time scales (most often three scales, from note-level up to a 10-second phrase). This is partially practical, as the HMM model-comparing process works best when the dictionary of gestures exist on the same time scale, but also because a note-level sonic gesture is musically very different from a phrase-level one, and needs to be understood accordingly. The process of traversing the different times scales is quite general, and can be extended to more layers: an onset (i.e. detection of an event) triggers the recognition of the smallest possible recognition level; if the accumulated confidence does not pass a given threshold within an adaptive time frame (determined by the average length of the gestural dictionary) then the system begins to try to make sense of the event as a “higher level” of gesture. This further allows for only one recognition process to be enacted at any given time, reducing the overall CPU load.

Whether or not a gesture is perfectly recognized, if the system detects a certain level of similarity, it can make decisions as to what output processing should occur. This output is a product of the mapping defined by the user, the instantaneous probability values and the smoothed confidence value. The overall musical output also takes into consideration the global sonic texture, which is a parallel listening process.

6. LISTENING TO SONIC TEXTURE

We also endow our system with the ability to listen to the global auditory scene in regards to its textural quality. Following the musical analysis presented in [20], we consider texture in our work as any sustained sonic phenomenon that possesses global stationarity (of amplitude and frequency content) while having local micro-variations. There is no clear separation between timbral events and textural sustain on one hand, but also between the local temporal variation or grain [6] of a sound texture and larger-scale modulations such as tremolo or vibrato. These phenomena differ along a continuum in regards to separability of temporal fluctuation. Our work explores the possibility of creating a machine listening system that creates a partition of the incoming audio in an adaptive and signal-driven way to listen along this continuum. To this end, we create our system based on the nonlinear time-frequency analysis technique called Empirical Mode Decomposition (EMD) [23], which separates a signal into a set of (not necessarily orthogonal) intrinsic mode functions (IMF) that are created not by differing spectral content per se (though this weakly is a by-product), but by different levels of temporal modulation. The technique is aimed at factoring out all components by virtue of their similar amplitude/frequency modulated content. We leave the details of our analysis to [20], but after separating the signal into this set of IMFs, these are clustered by virtue of their signal properties including the use of a roughness algorithm based on the work of [15], an analysis of envelope modulation and a measure that the first author has created known as temporal fine structure, defined on each signal block $L$ of size $N$ as

$$TFS(L) = \frac{\sum_{i=L+1}^{L+N} (x_p[i] - \frac{x_p[i]+x_p[i+1]+x_p[i+2]}{3})^2}{RMS(L)}$$  

where $RMS$ is the root mean square for the given signal block and $x_p$ is the instantaneous power-amplitude of the signal. After separating out the time-frequency IMF components and extracting these features – chosen precisely because they have proven effective in separating out qualitatively different types of signal grain [20] – we apply an expectation-maximization (EM) algorithm, which learns the dynamics of these different “spectral” regions in EMD-space (i.e. groupings based on type of temporal modulatory structure), and further detects the boundary between textural space. That is, when a global texture changes to a new scene, the algorithm notes this boundary. The full details of our approach are presented in [22], but the key idea is that we train the system on a set of exemplary musical textures (i.e.
recordings of our music that we characterize as complete textural sustain. The system analyzes the input audio and classifies this in terms of number and quality of each textural sub-group. It cannot be known when a texture has changed until after the fact, but we make use of our EM algorithm – the “E part” which is based on a Kalman filter – in order to maintain a running prediction of a potential change in texture on the “receding horizon”.

Therefore, our global texture-listening system is centered around listening to the sound from all players, deciding on the number and type of “sub-textures” the music is comprised of (e.g. long beating modulations, white noise and dust noise), and giving a short-term prediction of when a change in texture will occur based on this. Currently, we analyze the number and type of scenes in real-time, utilizing externals we have written in Max for analysis including the EMD technique, as pictured in figure 4. The example patch shows a mixture of two sine waves and white noise. Note how the external is able to separate the three components. In contrast to linear techniques such as the STFT, when the sine waves drift closer and cause beating, this modulation is identified as a separate IMF, much as a listener is able to detect this phenomena as distinct from the tones proper. We use this global texture-listening system as complement to the sonic gesture analysis, with a memory and learning that is updated in a way that is heavily based on current musical information.

7. LEARNING, EPISODIC MEMORY AND DYNAMIC ATTENTION

With the ability to segment and recognize individual sonic gestures as well as the overall musical texture in place, we believe our design allows for a machine listening that is attentive to changing gesture/texture content, as well as focal and global modes of listening that are particularly relevant in free improvisation as well as electroacoustic music. However, there is still the need to balance attention between these modes, to learn from the environment, and to create some form of memory that is engaging as a fellow improviser. The first issue is very much an open question for us, and we currently address this by mapping the output from each into low level parameters or higher meta-parameters for gesture recognition, and into high-level state changes for texture recognition.

Meanwhile, the issue of learning is addressed differently for these two modes of listening. The learning of sonic gestures is an adaptive process that mimics the EIS and GREIS paradigms of a moving window of delay-line memory (e.g. of 60 seconds in length). We treat this buffer as the system’s short-term episodic memory from which we learn new gestures. For this on-line learning, we segment the entire delay line buffer using the same onset analysis, and treat each potential event as a multidimensional feature vector. After building a K-Nearest Neighbor (KNN) tree, we define a novel gesture using one of two methods:

1) If an event that is sufficiently far in sound feature space from the other members of the known database is discovered, this is included in the gestural lexicon. If a max threshold is reached, members that are closer to the centroid of the “gesture space” are thrown away.

2) After clustering the KNN tree, the gestures that are closest to the centroid of each cluster define a new gesture space.

Using this method of learning, we either privilege recent performance memory completely as the relevant set of articulations to compare against, or we mix new gestures with those that we define a priori as being important. By contrast, we train the system on a set of sonic textural scenes a priori. This is necessary due to the nature of the learning algorithm, but also we have found this to be sufficient given the global and generalized nature of the listening task. Both of these parallel listening tasks are involved with the transition from short-term episodic to semantic memory, and define the immediate musical context.

The final issue is that of longer-term episodic memory. For this, we currently leverage the factor oracle pattern learning algorithm from the OMAX system [1]. This work is focused on learning a sequence of audio-level or MIDI events in order to learn stylistic trends without any assumption of style beforehand. One powerful aspect of this algorithm is that it allows for a succession of similar audio events to be identified as a single unit, with a precise time-stamp, and to reference past material that has structural similarity in regards to a left-to-right temporal sequence of audio events. In our approach we build such a structure over the duration of a performance, based on spectral features. The features related to gesture and texture are stored in parallel for each distinct sound event, so that when the player’s input audio relates to the stored short-term memory, the system can de-
cide to act on the longer-term memory that is precisely time-stamped and structured by the factor oracle algorithm. This oracle tree structure can then be navigated by the system in order to re-present audio, reacting to the currently recognized musical context. We believe that recalling a set of transitions between gestures and between general textural states is an appropriate level of structure to apply in listening during our own music-making, and in many other forms of improvisation as well.

8. MUSICAL ACTIONS

A full discussion of the musical transformations of our agent is beyond the scope of this paper – instead we have focused on the listening, memory and reaction aspects of our work as well as real-time implementations. However the key component of musical actions that we design for are directed spontaneity, gestural transformations and scene-aware transformations. The first of these was discussed in [21], and is based on the use of probability and confidence from gesture recognition to drive the selection, crossover, mutation and rate of a genetic algorithm. The population space consists of low-level parameters related to granular synthesis as well as selection of routing in a delay matrix. The second set of actions can be thought of as a sonic gestural analysis/synthesis: rather than quasi-random modulations on audio in a delay-line buffer, we scan the system’s episodic memory and select audio for transformation based on its gestural contour. The type of transformation – granular, additive-based or spectral envelope modulation – and specific actions are a product of the input sonic gesture. Finally, the highest level of musical actions – including density of events, switching speed between actions, overall level – are determined by the relative strength and ratio of the gesture/texture recognition.

9. CONCLUSION AND FUTURE WORK

Though focused on the musical activities of Triple Point, we believe our approach to system design based on listening and dual awareness to both gesture and texture is highly relevant to human/machine free improvisation in general. We have found our combined analysis, recognition and generative output approach to be very satisfying in extending our musical practice. Future work includes an extension into real-time global texture recognition and prediction, as well as meaningful recognition of larger musical trajectories at the level of an entire improvisation.

10. REFERENCES


