ADVANCING EXPERT HUMAN-COMPUTER INTERACTION THROUGH MUSIC

Benjamin D. Smith *Guy E. Garnett*
University of Illinois at Urbana-Champaign
Illinois Informatics Institute, National Center for Supercomputing Applications

ABSTRACT

One of the most important challenges for computing over the next decade is discovering ways to augment and extend human control over ever more powerful, complex, and numerous devices and software systems. New high-dimensional input devices and control systems provide these affordances, but require extensive practice and learning on the part of the user. This paper describes a system created to leverage existing human expertise with a complex, highly dimensional interface, in the form of a trained violinist and violin. A machine listening model is employed to provide the musician and user with direct control over a complex simulation running on a high-performance computing system.

1. INTRODUCTION

As software systems and cloud computing provide more and more powerful, complex computing tools it becomes necessary to discover new ways of augmenting and extending human control while retaining human judgment and analysis of complex situations. Popular computer-human interfaces for personal computing, such as WIMP elements and hardware, are typically designed generically to support as many different users and uses as possible. Today, physical computer interactions are becoming increasingly multidimensional, as advances in technology and understanding encourage movement away from conventional human interface devices towards hands-free gestural controllers. Additionally, developments in physical computer interaction systems, such as contemporary game console controllers and the newest 3D imaging cameras, are enabling new, specialized modes of interaction. Although these systems may not be as general, versatile, or approachable, the rewards in terms of capabilities in specific domains can outweigh the requirements of expertise and loss of generality. Given a user's willingness to practice and learn the new interface, such a system can give the user new levels of control and power.

Physical devices with extreme learning curves have long been in use, and we take the musical instrument, the violin, as a first class example. While being notoriously difficult to learn, the violin presents vast possibilities for sound creation and musical expression in the hands of a practiced master, audibly and visibly demonstrating the human capacity for using complex interfaces successfully. The typical professional musician today will spend upwards of one thousand hours per annum developing and maintaining their proficiency on the instrument. We then ask: what could this level of dedication lead to if the target was a computational control interface?

Another dominant trend in computing is increasing power and capabilities for complex data processing, data mining, simulations, and visualizations. This is especially apparent in high performance computing (HPC), with the recent completion and activation of several petaFLOP (10^{18} floating-point-operations-per-second) capable systems. In such a setting a user with extensive practice and skill could become a virtuoso of the computer, capable of transforming and manipulating vast data sets in novel ways, enabling new modalities of operation and knowledge discovery.

To explore this possibility we describe a novel interface to enable high-dimensional, continuous system input and control, leveraging the very precise gestural manipulations of professional musicians. This new interface builds on machine listening techniques, applying unsupervised machine learning algorithms to develop a machine model of expert human music listening. This setup presupposes expert knowledge of a musical instrument on the part of the user, effectively transforming the acoustic instrument into a digital control device. The result is the concurrent, precise, and direct manipulation of many independent parameters of a complex data simulation running on a supercomputer, employing the violin as a sophisticated, multimodal, tangible interface.

2. MOTIVATION

As available computing power and network bandwidth increase it gradually becomes possible to transform practices of high-performance computing and scientific data processing from offline, batch oriented modes to real-time interactive data mining and simulation. Interactive high performance computing is a growing area of research, demanding new systems and solutions [6]. The extreme level of sophistication that these interactive supercomputer jobs can achieve suggests the need for highly dimensional, tangible or gestural control interfaces in order to fully exploit the available computing power. To date, little work has been published in this direction.

Interfaces based on expert use of existing physical devices, and which are worthy of extreme practice must add as little additional cognitive load as possible. Towards this end it is desirable to develop an automated understanding of the current practice (i.e. making music), identifying the preexisting repertoire of gestures. Machine learning algorithms have proven successful in this domain, categorizing musical and physical gestures with sufficient accuracy and flexibility [5, 6]. Additionally, some approaches enable tracking of gestural qualities and styles [4] built of more or less subtle nuances and variations in execution. These elements of expressivity contain additional information that can be leveraged as a control source, potentially giving the computer access to a type of emotional understanding of the input.

3. MUSICAL STRUCTURE

Despite the many parallels that exist between spoken language and music, each domain presents distinctly different problems for machine listening. While language is highly semantic and denotative, music is more referential and connotative. Composers and theorists of classical musics typically describe structures and developments in music through relational patterns in the musical material (such as the return of a melody or a modulation to a new tonal area). These relationships are understood to give meaning, especially when realized through the expressive performance of a trained musician.

In this context we generically understand "gesture" to denote "a movement of the body that contains information" [7], which may be actuated as sound through the use of a musical instrument. This aural product is understood as a sonic or musical gesture (a movement of sound that contains information). These gestures in turn develop meaning through their ordering in a musical work or improvisation and the inter-gestural relationships they present.

Gesture parsing is accomplished through the employ of an online, unsupervised machine learning algorithm, Adaptive Resonance Theory (ART, [2], applied to music [5, 9, 10]). The ART model is a competitive neural network, in which all the nodes of the network compete for the best fit, with the winner adapting and receiving reinforcement. Effectively, each input sample is compared to a library of category exemplars and a simple distance measure is employed to find the best fit. If the match is within a specified distance tolerance the known category incorporates the new input (a smoothed filter of the current state and the new input). However, if no category match is found then the new input is taken as the initial state for a new category. In this way the input feature space is sequentially partitioned and mapped into categories that are germane to the given input sequence, and the granularity of category distinction can be easily configured for a given input set. ART networks are typically trained on-the-fly without employing any preselected training data sets.

4. ARCHITECTURE

The platform employed for the prototype is a distributed system comprising a supercomputer in combination with consumer grade desktop computers (fig. 2). The Abe supercomputer, housed at the National Center for Supercomputing Applications (NCSA), was used for the simulation core. A Mac Pro desktop computer was used to process and analyze the sound input and display the state of the simulation.

The system being controlled by the musician in this prototype was an agent based flocking simulation, modeling behaviors observed in nature as exhibited by herding and schooling creatures [11]. The non-centralized, particle nature of the algorithm is a distinct computational advantage, making it readily parallelizable and scalable. Each simulated entity maintains its own state and behavioral coefficients and parameters, updating its properties based on its immediately proximal neighboring entities with no concurrent dependencies in the update functions. As more entities are added the complexity of the algorithm proceeds as O(n^2), but the update calculations are shared uniformly across all available CPUs. [8] found unoptimized performance on the order of 50 updates per second with 10,000 entities on 128 CPUs, which was sufficient for proof of concept. We refer the reader to [8] for further details of the simulation implementation.

The data set from the simulation is rendered as a projected 3D visualization, displaying the flocking entities as uniform abstract shapes, rotating and moving dynamically in a virtual environment. Displaying the modeled behavior of the entities in a fashion analogous to flocking creatures observed in the physical world is a natural mapping given the nature of the simulation algorithm.

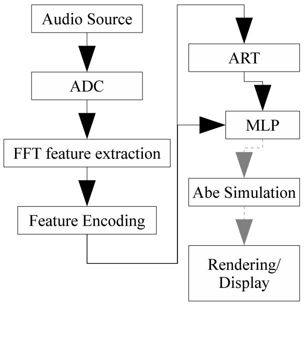

Figure 2: System architecture overview (gray stages indicate network transmissions).

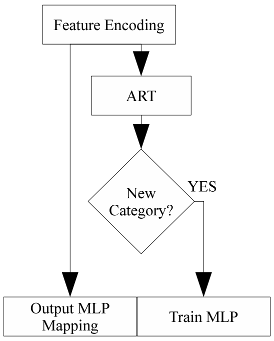

Figure 3: Detail of machine learning stages from fig. 2.

4.1. Sound Analysis

We take as our guiding premise the exploitation of existing skills possessed by trained musicians (as proposed by [7]). In order to fully leverage the sonic nuances presented by a given musician the computer must listen to particular components of the sound, and thus understand sound and music in a fashion that parallels human audition, especially that of the particular user.

Audio input from the violin is first passed through an analog-to-digital converter (ADC, see fig. 2), and then a Fast-Fourier Transform is performed, which enables the extraction of the target audio descriptors. These are encoded into a modeled short-term memory which is used as the input for the ART. The ART dynamically trains a multi-layer perceptron (MLP) neural network in order to affect the actual non-linear mapping process [10](shown in fig. 3). When the ART detects a new category it retrains the MLP in parallel and on-the-fly, allowing the MLP to continue mapping inputs to outputs uninterrupted.

The following descriptors are currently implemented in our system: Loudness, Dynamic movement, Spectral centroid, Noisiness, Pitch, Interval, Rate of pulse, and Temporal movement. Once sampled, the descriptors are assembled in a feature vector which is used as the input for the ART network.

Pitch and interval are treated with an additional step due to their privileged nature in defining musical shape. Western classical music typically employs a twelve tone equal-tempered scale for the mapping of the pitch-frequency space. This scale pattern is based on the notion of octave equivalence classes (i.e. every doubling of the pitch frequency is under- stood as the same scale degree but with different pitch "register"), employing twelve tokens to represent the twelve distinct scale degrees. While the pitch classes may be ordered in an ascending sequence their perceptual relationships are much more complex, serving to define notions of harmony, consonance, and dissonance independently from their order position.

The perception of melody is strongly based on the temporal order of pitches and the derived interval structure. To enable the comparison of melodic sequences we transform the input pitch class sequence into a recency vector through the process of spatial encoding [5]. This turns an ordered sequence of tokens into a regularly dimensioned vector where the position of the token in the sequence is encoded as a relative weight. The same spatial encoding is also applied to the interval sequence.

The ART layer now compares this input vector to known exemplars, describing the input in terms of match quality. While a single categorization is possible, we believe the match strength with all known exemplars more closely mimics human understanding of music. The resonance with each category provides a rich description of the current input and its structural location in the musical piece. For example, as a melody is repeated at the start of a piece it becomes a defined musical entity, for a given listener. Later any returns of that melody can be understood by the listener. Also, any variations in the returning melody can be described as well, relating the new modifications to other melodies that were exposed during the course of the work. While concrete, semantic understanding is not possible through the ART, access to the relative comparative value of melodic fragments is provided, paralleling some aspects of human understanding of music.

5. DISCUSSION

Testing of this system was conducted with one of the authors as a participatory designer and target expert user. Evaluation was carried out continuously during design and implementation, ensuring that the system met targets of usability, functionality, and accessibility to a trained musician. The resulting prototype enables clear demonstration of the musical control scheme, allowing the musician to make easily perceptive manipulations of the data simulation.

While the ultimate goal is a transparent leveraging of musical playing, the current system places a not insignificant level of cognitive load on the user. Musicians, especially when improvising, privilege their intuitions and a sense of emotional communication [1]. However, operating our system requires a different mental orientation that forces the musical user to form a very conscious memory of what they played during the session. Unlike most user interfaces designed today our system does not have a preconception of what input commands will be employed (as long as they are sonic in nature). This requires that the user define and learn their own control scheme. Since the ART network is learning on-the-fly, returning to the same musical place (in the sonic feature space) is necessary to reproduce the same outcome. On many occasions this proved difficult for the user, who struggled to recall their earlier improvisation.

Most of the observed difficulties in the use of the system centered around issues of intent and conceptual mapping. The complexity and adaptivity of the machine learning requires significant cognitive resources on the part of the user, who typically desires to exert precise control over the simulation. However, this is dependent on a precise understanding of the analyzed inputs (pitch, brightness, noisiness, etc.) which often are not heard focally by the player. This is an inherent challenge of the design, as the system is intended to respond intelligently and naturally, without requiring a strong cognitive model on the player's part. When the musician was able to 'let go' and focus on the music, creating a compelling sequence of sonic events, the control of the system become much more facile and transparent. This typically lead to the most rewarding experiences.

The primary user found our system promising and exciting overall. The above difficulties not withstanding, we were able to afford control of a complex, dynamic simulation for an expertly trained musician. Even failure in the control was mostly undetectable by any except the user, who can recognize a mismatch between intent and result. Additionally, this system leverages the intrinsic value and reward found in the actions employed in this control (i.e. making music), which could encourage sustained use of such designs.

While the information extraction provided by the ART network is very promising, employing this data in an appropriate and meaningful fashion remains a challenge. Ideally this mapping will be constructed dynamically and intelligently, serving to reduce the aforementioned cognitive load by making apparently natural or intuitive choices. This might be accomplished by associating some sense of affect with a given input mapping, such as by tying bold, heroic melodies to strong, quick changes in the simulation space, and mapping gentle, melancholic melodies to subtle, continuous movements (of course these mappings would be crafted uniquely for each given user, as descriptions of musical affect can vary dramatically from individual to individual).

We have shown that a complex data simulation, based on scientific observations, running on a supercomputer, can be dynamically controlled in real-time by a trained musician, demonstrating the potential for highly specialized, multidimensional expert interfaces. The prototype evaluates well in terms of control precision and rate of interaction, but demands a high level of extra-musical awareness and focus. Someday systems affording richly expert control will be prevalent and anticipating these situations can lead to the design of the best natural interactions possible, allowing users to learn new interfaces in the same way a violinist learns to play the violin.

6. REFERENCES

[1] Blum, D., and Quartet, G. *The art of quartet playing: The Guarneri quartet in conversation with David Blum*. Cornell U. Pr, 1987.

[2] Carpenter, G. A., Grossberg, S., and Rosen, D. B. Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks 4* (1991), 759–771.

[3] Dannenberg, R. B., Thom, B., and Watson, D. A machine learning approach to musical style recognition. In *Proc. International Computer Music Conference* (1997), 344–347.

[4] Fiebrink, R., Cook, P. R., and Trueman, D. Play-along mapping of musical controllers. In *Proc. of the International Computer Music Conference* (2009).

[5] Gjerdingen, R. O. Categorization of musical patterns by self-organizing neuronlike networks. *Musical Perception* (1990).

[6] Kurtenbach, G., and Hulteen, E. *Gestures in human-computer communication*. Addison Wesley, 1990.

[7] Pietrowicz, M., McGrath, R., Smith, B., and Garnett, G. Transforming human interaction with virtual worlds. *Workshop on Computational Creativity Support at CHI 2009*, April 4, Boston.

[8] Smith, B., and Garnett, G. High performance computing for music. In *Proc. of the International Computer Music Conference* (2011).

[9] Smith, B., and Garnett, G. The self-supervising machine. In *Proc. of the International Conference on New Interfaces for Musical Expression* (NIME) (2011).

[10] Smith, B., and Garnett, G. Machine listening: Acoustic interface with art. In *Proc. of the International Conference on Intelligent User Interfaces* (IUI) (2012).

[11] Spector, L., Klein, J., Perry, C., and Feinstein, M. Emergence of collective behavior in evolving populations of flying agents. *Genetic Programming and Evolvable Machines 6*, 1 (2005), 111–125.