

Machine Listening: Acoustic Interface with ART

Benjamin D. Smith

University of Illinois at Urbana-Champaign
bdsmith3@illinois.edu

Guy E. Garnett

University of Illinois at Urbana-Champaign,
Illinois Informatics Institute
garnett@illinois.edu

ABSTRACT

Recent developments in machine listening present opportunities for innovative new paradigms for computer-human interaction. Voice recognition systems demonstrate a typical approach that conforms to event oriented control models. However, acoustic sound is continuous, and highly dimensional, presenting a rich medium for computer interaction. Unsupervised machine learning models present great potential for real-time machine listening and understanding of audio and sound data. We propose a method for harnessing unsupervised machine learning algorithms, Adaptive Resonance Theory specifically, in order to inform machine listening, build musical context information, and drive real-time interactive performance systems. We present the design and evaluation of this model leveraging the expertise of trained, improvising musicians.

Author Keywords

Artificial intelligence, unsupervised machine learning, machine listening, music, adaptive resonance theory.

ACM Classification Keywords

H.5.m [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

General Terms

Algorithms, Human Factors, Performance, Theory

INTRODUCTION

On-line machine learning (ML) continues to gain increasing application as the problems of real-time interactive system control become more and more approachable and better understood [4]. In particular, the promise of the transparent union of live performer and complex multi-media performance through intelligent interfaces is alluring to many, and has been the focus of continuing research over the last several decades. With the development of on-line ML algorithms and the computing power required for real-time operation, such performances are rapidly becoming reality. Yet, most musical ML applications developed to-date use pre-trained neural networks or other supervised models, typically requiring extensive pre-performance training.

This presents an unusual challenge to the improvising performer, as once the training is complete the musician is constrained to the pre-selected material. Yet, what if the computer could effectively participate as an intelligent partner, listening to the musical development and making informed choices and mappings *as the improvisation unfolds*? Musical context, the consciously and unconsciously perceived relationships between musical events within a given work, informs human perception of a piece through the formation of expectations [9]. Building a similar model of context and expectations for a computer promises parallel comprehension and the potential for truly intelligent interactions and responses.

In order to privilege the creativity and intuition of an improvising performer qua user we propose a system that listens as a trained human might, extracting relevant information from the music as it unfolds. The resulting data in turn can be mapped to the control inputs of any interactive system. This model employs an on-line, unsupervised machine learning implementation, based on Adaptive Resonance Theory (ART, [1]) algorithms to efficiently process any musical improvisation or input.

BACKGROUND

The applicability of ML methods to problems in interactive musical performance is evidenced by the number and variety of applications and cases (see for example [14, 15]). Recent systems, such as the work of [4], focus both on “real-time” training, in order to better match the musician’s work process, as well as the use of ML to discover new musical expressions. Rather than attempt to exactly duplicate preconceived mappings they encourage the exploration of unexpected results stemming from active training during a performance.

However, all of these implementations require that the user define both their input material as well as the desired outputs before use, relying heavily on the acuity of the users in defining appropriate control spaces. Yet, if the user is inspired to take the musical improvisation in a new direction the connection to the computer becomes challenged, as the content of the music moves away from the domain that the system was trained for.

The potential to capture the informational content without pre-performance training is presented in unsupervised, on-line ML models, which have seen virtually no application in interactive performance to date. These models allow the system to discover classifications and find groupings and patterns across inputs based on relationships inherent in the data, rather than training on preconceived, external knowledge of the inputs. Effectively, the computer is allowed to build its

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IUI’12, February 14–17, 2012, Lisbon, Portugal.

Copyright 2012 ACM 978-1-4503-1048-2/12/02...\$10.00.

own interpretation of the musical work, listening in a fashion analogous to the human listener [2]. Unsupervised methods lend themselves naturally to sequential input processing, which is a prominent aspect of human perception models, rather than all-at-once training.

By giving the computer the capability to discern musical priorities and relationships on-the-fly, mappings can be constructed that serve to expose the musical development (and potentially emotional cues [7]) of the improvisation. This minimizes user cognitive load and allows the improviser to focus on playing and creating a compelling musical work, speaking a language of melodic and textural relationships without worrying about the computational details. The capabilities of unsupervised learning to analyze musical material is shown by [5] and [12]. While our work focuses on interactive computer music, the same models can be applied to other intelligent, interactive computer control systems with high dimensional, real-time input.

ART

Basing our machine listening implementation on contemporary theories of human perception and audition [10], we employ a multi-layer design to parse melodic input streams. The human model comprises both a short-term memory (STM), to retain the last several seconds of sound input, and a long-term memory (LTM); grouping, relating, and categorizing inputs for later retrieval. Both of these memory functions feed into higher level processes that discern patterns, extract knowledge, and make decisions.

Spatial encoding [3], a method employed in natural language parsing, is employed to characterize sequential input data, transforming melodic context into a twelve-dimensional spatial representation (one dimension for each pitch-class in 12-tone equal-tempered tuning). Spatial encoding employs an unordered single layer Neural Network (NN), with one node for every token in the input set. When a token is presented to the NN the associated node is fully activated and the activity of the network x is attenuated by a small amount α : $x_t = \alpha x_{t-1}$. This model preserves the ordering of inputs by creating a vector encoding, where α may be variably set to control the length of the “memory” (typically to retain 5-9 tokens, based on Miller’s Law [8]). This system is used in [5], for example, to successfully recognize formal relationships in early Mozart piano works.

The biggest limitations of the standard spatial encoding are the inability to describe repetitions and the generic nature of the activation. Two sequences, such as AABC and BABC, can produce the same encoded vector, yet in musical contexts sound distinctly different. Additionally, it has been shown that the human ear weights sonic input differently based on accents of agogic, dynamic, and metric inflection [11]. However, spatial encoding treats all tokens equally.

We propose a modification to the spatial encoding NN to dynamically adjust the node activation levels based on the perceived attention a particular token warrants, which we term Dynamic Encoding. The update rate is now set at a rapid sample rate (~ 20 hz) and the activation level is set nominally

to increment by a small value (such as 0.05, to avoid over-saturating the network). In practice this gives proportionally more weight (a more active node) to longer notes, while short notes only appear in the network for a few update cycles. Other types of accents are incorporated by increasing and decreasing the activation value appropriately (i.e. louder notes get more activation and softer notes receive less).

The activation level (l) for dynamic accents is calculated based on the amplitude of the current sample (a_t) and a running average (of length k , typically 16) of the amplitudes of previous samples.

$$l = \frac{a_t k}{\sum_{i=1}^k a_{t-i}} \quad (1)$$

Thus $l = 1$ for equal amplitude notes, $l > 1$ for louder notes, and $l < 1$ for quieter notes. The activation level is then scaled to account for the rapid update rate (by 0.05 in our tests below).

Finally, the STM decay function has a significant impact on the shape derived from the melodic input. The standard model is a linear function, $x = x_{t-1} - b$, or exponential. We also consider a sigmoid curve, $x = x_{t-1}^c$, where c controls the effective temporal length of the memory module.

Long-Term Memory

Once the STM is created, the resulting vectors are fed to an ART module for classification. The ART is a competitive NN that trains in an unsupervised manner, discovering categories based on an ordered sequence of input vectors. Different orderings of the same data will produce divergent classifications and while this lack of consistency is typically seen as a limitation, here it is an asset. Just as a human listener relates later melodic development (such as a sonata-form development section, or recapitulation) to earlier auditions (i.e. the exposition) based on the meaningful ordering of music, so does the ART algorithm. The details of the ART algorithm are described at length by [1]. Two parameters of significance are the category size limit, or “vigilance” parameter (p), and the learning rate (β), which allows the network to both train new inputs immediately and still adapt gradually, retaining the identity of older categories. Setting the learning rate high causes categories to fully incorporate new inputs while setting it low causes the categories to adjust slowly, settling into an average area of the category space.

EVALUATION

We describe the ART analysis of an improvised solo by violinist Jean-Luc Ponty, “No Strings Attached,” from the album *Le Voyage*. This recording was chosen due to its commercial availability, evident virtuosity, and range of harmonic and melodic content. The digitization involved running the recording through a pitch tracker to produce a stream of paired pitch and amplitude values (one sample approximately every 50 milliseconds for 6892 samples total). This ensured comparability between tests by removing any variability in the pitch reduction process. The recording selected is $\sim 5' 56''$ in duration, consisting of a gradual exposition of material with

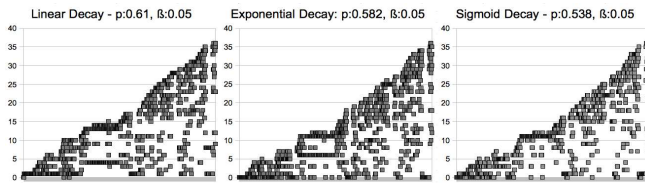


Figure 1. Categorizations with Spatial Encodings (linear, exponential, sigmoid) of *No Strings Attached* (category ID vs. time).

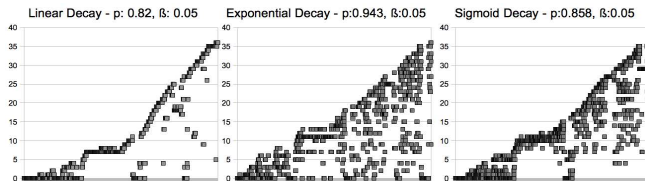


Figure 2. Categorizations with Dynamic Encodings (linear, exponential, sigmoid) of *No Strings Attached* (category ID vs. time).

several sections dominated by looped drones and rhythmic ostinati. We focus solely on melodic material, incorporating rhythmic duration and dynamics, as this is the primary index for pattern identification in human musical cognition [6].

For comparison we employ six STM models: spatial encoding and dynamic encoding each with three decay models (linear, exponential, and sigmoid). Decay rates were set so that each model effectively retains 7 ± 2 inputs at each update. Using slower decay rates tends towards gradually increasing category creation (as more context information is retained), while quicker decay rates produce far fewer categories (only twelve at the limit). Both of these extremes lead to problems building a useful understanding of musical context.

ART Classification

Next we feed each of these STM outputs into ART modules to produce categorizations and find patterns in the feature vectors. Fig. 1 and 2 show the categories generated over the course of the recording for each STM setting noted above (ART vigilance, p , and learning rates, β , are set to ensure a comparable number of categories—36—are identified for each encoding model). Note that the vigilance settings need to vary greatly to ensure the same total category creation for the very different STMs. This further emphasizes the impact of each encoding model on the characterization of the feature data.

The unfolding of the work produces gradual category creation for all encodings, which is appropriate to this free improvisation. The impact of the different encoding schemes can be seen. Some encodings, the dynamic encoding with linear decay for example, suffer from overfitting, creating new categories with few returns to old material. Others, such as the spatial encoding with exponential decay, are more reductive, showing apparently random repetition without sufficiently accounting for context. However, all indicate a move to a second musical section near the middle (esp. the dynamic sigmoid encoding), and hint at brief returns to the initial mate-

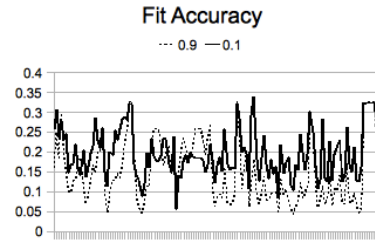


Figure 3. Accuracy of fit, *No Strings Attached*, spatial encoding, sigmoid decay, $c = 2.0$, $p = 0.9$.

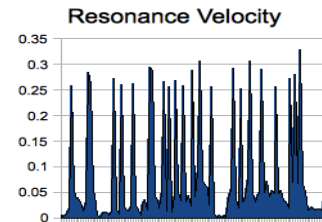


Figure 4. Euclidean distance between resonance vectors, *No Strings Attached*, 2' 53" to the end. [Dynamic encoding, sigmoid decay, $c = 1.10409$, $p = 0.858$, $\beta = 0.05$]

rial in the middle and near the end (category 0 at the bottom of each graph).

The vigilance parameter was set in order to produce a comparable number of categories. The effect of the vigilance on the category generation rate is roughly exponential in nature, resulting in ~ 50 at $p = 0.9$, ~ 180 at 0.75, and ~ 1100 at 0.95 (for our chosen example with a spatial encoding and sigmoid decay). The learning rate parameter indirectly affects the rate of category creation (varying from a total of 8 categories when $\beta = 0.01$, with a dynamic, sigmoid decay, $p = 0.9$, to 40 when $\beta = 1$). Distinctly different inputs are still recognized as such, regardless of the learning rate. The learning rate parameter is also visibly at work in the “accuracy” measure of the input classification. This is calculated by taking a simple Euclidean distance measure for each input vector to the center of the matched category (fig. 4). The higher learning rate (0.9) results in more accurate fits generally while the lower rate (0.1) is looser, adapting to new inputs more slowly. The tradeoff is between the nature of the information sought: rapid, accurate categories, or slower, more general categories.

Resonance

Given the STM and LTM above, the extraction of information regarding musical relationships is now possible. Sectional movement can be identified by large movements in encoded feature space, resulting from changes in pitch material (modulations and transpositions). As fig. 4 shows, the distance between successive resonance vectors varies dramatically, indicating movement between different musical sections. Short periods of relative stasis are followed by sudden shifts to different areas of the category space, denoting the same shift in the source musical material.

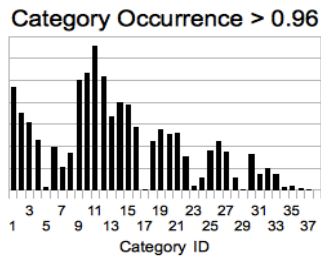


Figure 5. Total occurrences of each category at a resonance of 0.96 or greater. *No Strings Attached*, Dynamic Encoding with sigmoid decay, $c = 1.10409$, $p = 0.858$, $\beta = 0.05$.

Additionally, counting the number of matches for each category can be translated as a measure of the importance of each category. Taking the resonance vectors, and thresholding the resonances of each category observation, it is possible to compute a relative import for each category. Fig. 5 shows the ratio of category resonances observed with a value over 0.96. Certain categories appear with much greater frequency, while others (categories 5, 17, 23, 29, and 37, for example) only appear rarely. Using this knowledge the computer could decide to map significant outputs to strongly recurring categories, or the converse, highlighting the fleeting, unexpected categories.

Analyzing the resonance vector for these types of shifts and changes is the goal of this process. They are readily apparent to the trained human ear, and through statistical processing they can be appreciated by the computer “ear” as well.

FUTURE DIRECTIONS

The STM and LTM models set forth above provide an analysis of a live improvisation, extracting context information through the categorization of melodic feature data and the measure of relationships and movements within the melodic feature space. At the moment, the task of mapping this data to an interactive performance system remains predominantly in the hands of the technician or composer. Automating this process will require additional analyses of the fit data (both the resonance vectors and accuracy) over time, locating significant categories, tracking the movements in feature space, and mapping their inputs to desirable outputs. Both neural-network type dynamic mapping systems as well as ARTMAP may lead to viable solutions. This promises rich mapping possibilities that might match the perceptive complexity of the musical improvisation.

The models discussed thus far are currently being further evaluated and employed in a suite of pieces, transforming the musical improvisation of a live violinist into animations and audio through a granular synthesis engine [13].

ACKNOWLEDGEMENTS

The National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Institute for Advanced Computing Applications and Technology, and the Illinois Informatics Institute.

REFERENCES

1. 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.
2. Collins, N. *Towards Autonomous Agents for Live Computer Music: Realtime Machine Listening and Interactive Music Systems*. PhD thesis, University of Cambridge, Cambridge, UK, 2006.
3. Davis, C. J., and Bowers, J. S. Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. *Journal of Experimental Psychology: Human Perception and Performance* 32, 3 (2006), 535–557.
4. Fiebrink, R., Cook, P. R., and Trueman, D. Play-along mapping of musical controllers. In *Proceedings of the International Computer Music Conference* (2009).
5. Gjerdingen, R. O. Categorization of musical patterns by self-organizing neuronlike networks. *Musical Perception* (1990).
6. Hébert, S., and Peretz, I. Recognition of music in long-term memory: are melodic and temporal patterns equal partners? *Memory & Cognition* 4 (1997), 518–533.
7. Krumhansl, C. An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 51, 4 (1997), 336.
8. Miller George, A. The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychological Review* 63 (1956), 81–97.
9. Pearce, M., and Wiggins, G. Expectation in melody: The influence of context and learning. *Music Perception* (2006), 377–405.
10. Peretz, I., and Zatorre, R. J. Brain organization for music processing. *Annual Reviews Psychology*, 56 (2005), 89–114.
11. Pfordresher, P. The role of melodic and rhythmic accents in musical structure. *Music Perception* (2003), 431–464.
12. Piat, F. G. P. *Artist: Adaptive resonance theory to internalize the structure of tonality*. PhD in human development and communication sciences, University of Texas, Dallas, Aug. 1999.
13. Smith, B., and Garnett, G. The self-supervising machine. In *New Interfaces for Musical Expression* (2011).
14. Thom, B. Interactive improvisational music companionship: a user-modeling approach. *User Modeling and User-Adapted Interaction* 13 (2003), 133–177.
15. Wessel, D. Connectionist models for musical control of nonlinear dynamical systems. *The Journal of the Acoustical Society of America* 92, 4 (Oct. 1992), 2402.