Preprint. To be published: Computers in Entertainment 14(4), 2016

Towards Machine Musicians Who Have Listened to More Music Than Us: Audio Database led Algorithmic Criticism for Automatic Composition and Live Concert Systems

NICK COLLINS, Durham University

Databases of audio can form the basis for new algorithmic critic systems, applying techniques from the growing field of music information retrieval (MIR) to meta-creation in algorithmic composition and interactive music systems. In this article, case studies are described where critics are derived from larger audio corpora. In the first scenario, the target music is electronic art music, and two corpuses are used to train model parameters, then compared with each other and against further controls, in assessing novel electronic music composed by a separate program. In the second scenario, a "real world" application is described, where a "jury" of three deliberately and individually biased algorithmic music critics judged the winner of a dubstep remix competition. The third scenario is a live tool for automated in-concert criticism, based on the limited situation of comparing an improvising pianists' playing to that of Keith Jarrett; the technology overlaps that described in the other systems, though now deployed in realtime. Alongside description and analysis of these systems, the wider possibilities and implications are discussed.

• Applied computing ~ Arts and humanities ~ Sound and music computing • Information retrieval ~ Multimedia and multimodal retrieval ~ Music retrieval

Additional Key Words and Phrases: Musical machine listening, machine learning, music corpus study, algorithmic composition, live computer music, algorithmic critic, computational aesthetic function

1. INTRODUCTION

Human musicians live out their musical lives according to a panoply of influences, against a backdrop of manic cross-cultural musical activity. The mass of music released every day exceeds anything it would be plausible for an individual to listen to in a lifetime, and is substantially modulated by cultural filters such as the reception bias of an individual's information seeking behaviour, or the advertising budgets of music companies and other socio-economic factors. Encultured musical development takes many years, and expert musicians are likely to have studied an instrument for at least ten thousand hours. For machine systems to stand a chance of engaging in creation on human terms, they must inevitably treat the years of training and cultural absorption involved in human musicianship. A machine learning perspective is unavoidable.

Whilst there are precedents in statistical learning over symbolic music databases for algorithmic composition (for example, David Cope's work [Cope 2001]), less work in musical meta-creation has actively utilised audio databases, the raw and natural material of human musical culture. The big data problem of musical audio has been most obviously tackled by pursuits in audio content led Music Information Retrieval (MIR) [Casey et al. 2008], though MIR systems themselves are typically deployed offline on large audio collections, and rarely attempt realtime work and novel content creation. Audio content based MIR necessarily engages with the core computer music problem of musical machine listening, and current generation MIR technologies can be critiqued for the level of musical listening they actually embody [Sturm 2014].

In this article, the potential of large audio databases for computational critics and aesthetic functions is examined. Computational aesthetics has taken the critic function as the core of emulation of artistic activity [Stiny and Gips 1978; Galanter 2012]. A number of case studies are discussed herein, where critics for non-realtime algorithmic composition, and for realtime interactive music systems, have been constructed from large audio databases. In particular, UbuWeb and empreintes DIGITALes electronic music corpuses (both multi-day audio corpuses), electronic dance music audio files with a focus of dubstep, as well as a database of Keith Jarrett piano music, are used to derive critic functions that can act on newly observed material. Judgments may form part of a simulated iterative compositional design process, model an act of music criticism or competition judgement, or even be live feedback to a performer.

This article outlines a set of experiments in new research on critic functions derived from large audio databases at a scale hitherto unexploited. The originality is the move towards "more audio than it is reasonable for a human being to listen to" where the training database is significantly scaled up, and in the entrusting of critical tasks to machine proxy; such musical cultural modelling and involvement is central to future pursuits in MuMe.

After a discussion of the historical, technical and aesthetic status of algorithmic critics in section 2, the article details the construction and comparison of two critics for electronic art music in section 3, as well as a contrasting project in the automatic judging of a dubstep remix competition in section 4. Section 5 considers application for live in-concert automatic criticism, with an example based in piano improvisation. Discussion of the research progress, and the ramifications for future research and music making, conclude. The research undertaken here provides a feasibility study within current generation MIR-led machine listening and learning, with discussion of the technological and aesthetic potential.

2. ALGORITHMIC CRITIC TECHNOLOGY

Critical appraisal of music is essential to the deeply considered acts of composition and improvisation by musicians, as well as to the reception in culture of newly created music. Robert D. Schick, in his book *Classical Music Criticism*, locates the essential role of the critic as the timely analysis of musical evolution, corresponding to Oscar Thompson's 1934 exhortation to 'hold up a mirror' to the musical event [Schick 2013, p. 21]. Leonard Meyer [1973] sees the critic as explaining 'order already present in some work of art' (p.4) though criticism cannot cover all individual subjective reactions, but the application of an understanding of a wider corpus of music and pertinent music theory to a work in question: 'critical analysis uses the laws formulated by music theory–and, as we shall see, the normative categories of style analysis–in order to explain how and why the particular events within a specific composition are related to one another' [Meyer 1973, p.9]. Why did a creator take their particular choices to arrive at the work in question?

The term algorithmic critic has been popularised by Stephen Ramsay in the context of automating literary studies [Ramsay 2011]. Within the tradition of algorithmic composition and interactive music systems involving a generative component, a critic has a less refined cultural role than that described in the previous paragraph, but acts as a filter on permissible generated material [Fernández and Vico 2013; Compton et al. 2013; Machado et al. 2003; Cope 2001; Todd and Werner 1999; Rowe

1993]. This might be instantiated through a module of code, such as a fitness function in a genetic algorithm, or the set of tests for successful material to pass in a 'generate and test' paradigm. Spector and Alpern write of 'an artist construction system that takes as input a set of critical criteria and a case-base of past artworks' [Spector and Alpern 1994, p.4]; new works are created based on a model populated via the features of the case base, and assessed using the criteria provided. The paper is strongest in its recognition of the problems with formalization, identifying three primary issues:

1) Dead forms (algorithmic composition tends to be applied within well-established and historically distant styles)

2) Rules may lead to mediocrity 'it is not clear that adherence to the rules of a particular art form is a good indicator of aesthetic value; it might merely indicate inclusion in the genre, which might be compatible with aesthetic mediocrity' (p.4)

3) Criteria supplied may not generalize across artworks, 'many of which seem to resist the imposition of criteria upon which the art world can consense' (p.4)

We return to consider these issues in the later discussion section of this paper.

In the present study, much of the work is intimately tied to a mainstay of MIR, the similarity measure. Human responses to newly created music often revolve around statements of similarity to previously encountered music: 'It reminds me of Frank Zappa's mail order album guitar soloes crossed with Francesca Caccini's recitatives in her 1625 opera ...' A composer presenting a new work may feel enervated to be hauled back to earth by such presumptions of influence, or enjoy the new combinations and connections so evoked. The mass of prior art in culture is an inescapable backdrop to the act of composition. When machines are involved in making musical judgments, the volume of reference can potentially scale up to even more depressing or exciting a degree, depending on a creator's perspective. In the main, positive benefits of historical awareness and the potential for new "database music" may be hoped to dominate: A recent online study of musicians' attitudes to influence found no strong evidence for any 'anxiety of influence' and most musicians in the survey welcomed a sense of musical history and culture [Collins 2011].

Since the notion of similarity remains critical to any decision, musically meaningful descriptors are key to the integrity of such a method in the computational domain. It is impossible to claim current generation listening machines are on a par with human expert listeners [Sturm 2014], so that similarity measures are a weak point to be made explicit in research work. Human critical judgments may rest on parallel consideration of multiple attributes at a high level of musical understanding, and the weighting of these heard-out qualities, as well as the signal processing problem of extracting equivalent information, remains beyond the research frontier. Nonetheless, much can be explored around these issues using current generation feature extraction, and a strong feel for applications for future machine listeners explored.

MIR-inspired criticism has a small but fascinating literature. Brian Whitman and Dan Ellis [2004] provide a canonical study, describing an attempt to match up record reviews from two sources, the more musically feature oriented allmusic.com and the more polemical and off the wall pitchfork, to an audio database of 100 pieces, in order to form an automatic text writing critic for audio. Subsequent work by other authors has concentrated more on associations between text description and musical content [Turnbull et al. 2008; Bertin et al. 2010], or when considering music criticism, tackled text alone [Hu et al. 2005]. Studies have only rarely considered live music making: Jewell et al. [2010] exploit MIR techniques to query for similar audio to that of a fresh jazz improvisation, though on a limited scale considering a few entry phrases and a small corpus of material, and not in real-time. Nonetheless, the concern to track derivative or novel playing, and allow a musician user to compare themselves to a wider world of audio then their own immediate present, has a strong overlap with the motivations behind the current work.

3. ELECTROACOUSTIC ART MUSIC CRITICS UTILIZING UBUWEB AND EMPREINTES DIGITALES CORPUSES

For a test study in training critic functions from larger audio databases, two corpuses were explored. The first corpus is a set of historical electronic music, dating from 1937 to 2000 with a few gaps, available online from the art resource site UbuWeb (http://www.ubu.com/sound/electronic.html), and consisting of 476 files totaling 2.3 days of audio. The second is a collection of the first 120 releases, from 1990 on to 2013, of the respected Canadian electroacoustic music label empreintes DIGITALes (eD), purchased from the label for research purposes. The audio here covers 919 files, and 5.7 days total playing time. Such audio database sizes, large as they are, are not unusual these days in the MIR domain, though given commercial links, MIR typically treats popular music. The corpuses utilized for this research are highly specific to the electronic art music repertoire. Further, the task of novel algorithmic composition has not previously worked with such rich audio file databases.

Feature	Notes
Perceptual loudness	Utilizes an auditory model
Sensory dissonance	Following William Sethares
Spectral centroid	Brightness of sound
Average attack slope	Measured over the last two seconds
Spectral entropy	Entropy of momentary spectral distribution
Transient detection strength	As measured with an onset detection function
Event attack distribution statistics	Attack density (attacks per second), mean inter-onset interval (IOI) in a two second window, standard deviation of IOIs in a two second window
Beat histogram statistics	Entropy of beat histogram (entropy of distribution of energy recurrences at different tempi), metricity of beat histogram (how well a fundamental beat frequency explains the beat histogram, e.g. clarity of beat tracking)
Band-wise signal energy	Low, mid-range and high frequency energy

Figure 1 Features extracted.

Both corpora were analyzed with the same feature extraction front-end and model training. Fourteen features were extracted, as indicated in Table 1. Full formal definitions of all these features would go outside the scope of the article, but source code is available on request, and the extraction used the SCMIR analysis library for SuperCollider (http://composerprogrammer.com/code.html); the features mentioned here are directly based on machine listening capability available within this environment. The nature of these audio features is relatively low level, and does not include pitch information outside of a general measure of brightness, and three filter bands; this is a deliberate choice given the nature of the electronic sound material and the limitations of polyphonic pitch detection at the present time. Onset detection and beat tracking based extraction is included, however, as well as timbral measures such as the spectral centroid and attack slope (such features arise as important in perceptual studies). Following extraction for any individual audio file, all features were normalized with respect to maximum and minimum occurring values, within each given feature, across the whole corpus.

Having obtained feature trails at the rate of around 43 values per second, aggregate values were derived through texture windows formed based on mean, standard deviation, max, and min values. Windowing aggregated by two second windows with a hop size of one second. The original 43 values per second were not kept, as simply providing too much data to comfortably train models, and also because two second aggregation is truer to the short-term perceptual present of human working memory.

Model training is related to the critical task to be undertaken upon newly observed audio files (e.g., newly generated work to be assessed). Representative challenges might be:

- Measuring novelty versus the corpus
- Creating a year prediction function (chronological locator)
- Finding the closest matching pieces in the corpus (specific precursors)

though further tasks can be considered. Different machine learning algorithms are suitable, and can be compared in performance, with an eye to the pragmatics of the time required to train and test models. For example, a k nearest neighbors algorithm would be straight forward for finding matches, but scaling up efficiently for larger databases requires pre-calculation of special structures such as k-dimensional trees or approximate methods rather than exhaustive search [Slaney and Casey 2008]. Year prediction might utilize standard machine learning algorithms such as a neural net, support vector machine or Naïve Bayes, working by years or more realistically by ranges of years. Similarity measures can involve a number of metrics, and work from summary feature vectors, sequences (time series) of feature vectors, or tagged classes/integers representing such vectors [Casey and Slaney 2006]. This article concentrates on the first task of measuring novelty versus the corpus, as highly pertinent to computational creativity, though the others have been explored elsewhere [Collins 2015].

For this experiment, summary feature vectors for two second texture windows, one per second of each audio file, were obtained as above, and then subject to reduction to integer sequences via a k-Means clusterer (with k=20). The integer sequences were used to train a prediction by partial match variable order Markov model (specifically and technically, a 5th order PPM-AX "C" variant with escape but not exclusion, as per [Pearce and Wiggins 2004]). The justification for the use of PPM is supplied in a previous publication [Collins 2015]. Once trained, the model can observe an input integer sequence, and judge its unpredictability with an average log loss measure, defined over a sequence {x_i} of length N where the probability according to the model of the ith value x_i is $P(x_i)$ (given previous values up to the order of the model):

$$average\ \log loss\ (\{x_i\}) = \frac{1}{N}\sum_{i=0}^{N-1} -\log\left(P(x_i)\right)$$

The measure provides a proxy for novelty. That is, the harder the model finds it to predict which integer comes next in a sequence, on average, the less the sequence is explained by the trained model and hence the more novel this input. Logarithms are used to avoid numerical problems which otherwise arise when working directly with products of probabilities. Higher average log loss denotes a less predictable sequence since lower probabilities lead to larger negative logarithms. The PPM model shows a low log loss on its training data, and always higher for novel material.

Two models were created, one for the UbuWeb corpus, and one for the empreintes DIGITALes. A common global normalization was used (that derived from the complete eD corpus) for compatibility of feature vector values. The pieces from a given corpus could then be evaluated by the model created with the other corpus, and both models applied to novel material.

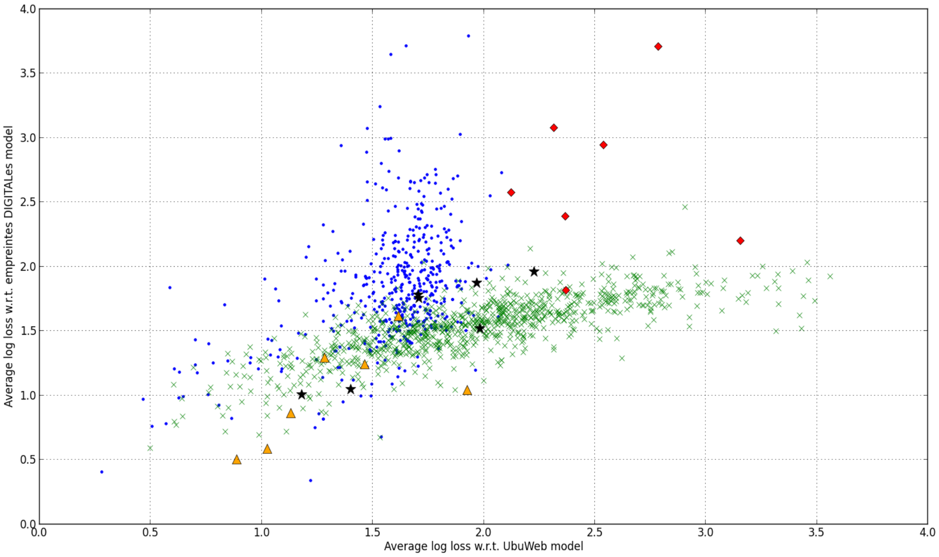

Figure 2 Plots of electronic music works with respect to the predictive power of two models; one model is derived from the UbuWeb corpus (x axis), and one from the empreintes DIGITALes (eD). Blue points are the 476 UbuWeb pieces, green 'x's are the eD pieces. Seven orange triangles denote seven Kraftwerk pieces; seven red diamonds denote seven algorithmically composed electroacoustic works; seven purple stars represent seven established electroacoustic works not in either the UbuWeb or eD corpuses.

Figure 1 plots the average log loss of model A (UbuWeb) on the x axis versus that of model B (eD) on the y, plotting all the 476 UbuWeb pieces, 919 empreintes DIGITALes pieces, and some further control audio files and algorithmically composed pieces of interest in this study. Points located closer towards the lower left corner can be interpreted as individual pieces well predicted by both models; position towards the upper right indicates works poorly predicted by either model. The points plotted by triangles correspond to seven Kraftwerk pieces (synth pop and electronic dance music trailblazer works dating from 1974 to 1981), and their position in the diagram shows that they are relatively well predicted by both models, though not always within the population mass of either. This predictability is related to the heightened repetition characteristic of the formative synth pop/EDM style, leading to a lower variety of transitions over time.

Points plotted by diamonds correspond to seven outputs of an automatic electroacoustic art music generation program. Generally falling outside of both corpuses and poorly predicted by their associated models, the implication is that the algorithmically composed works are neither electronic music classics (the UbuWeb corpus, in some sense), nor suitable for submission to the eD record label! There is one work which does overlap the tail of less predicted eD pieces from the eD corpus; if choosing one work to submit for consideration, this work would seem to match the aesthetic of the label best (within any trust of the machine listening and modeling assumptions underlying this decision). As a further control on such assertions, seven further pieces, this time established electronic music works in neither the UbuWeb nor eD corpuses (Trevor Wishart's *Imago* (2002) and six works from *New Music For Electronic & Recorded Media. Women In Electronic Music – 1977*, New World Records 2006) are plotted. Their general position is more central comparatively to the UbuWeb and eD art music works.

The plot doesn't indicate musical similarity with respect to musically meaningful attributes, but a level of explanation with respect to two models which notionally follow temporal changes in audio feature vector labels. Interestingly, limiting the variable order Markov model to a lower maximum order compromises the distinguishing power of the diagram, overlapping many more of the pieces from the two larger corpuses, and reducing the discrimination with respect to novel observed material. The higher order of 5 used here specializes the PPM models more to each corpus, though the existence of some overlap in the plot still indicates there is no over-specialisation and generalization is still possible. The lack of success in predicting the algorithmically composed works may seem to be an indication of originality, though the distance from both corpuses may also indicate going too far from established work in electroacoustic music. Indeed, previous evaluation of the algorithmic composition program involved [Collins 2012], which included multiple human evaluators, showed it to operate only at the level of an early undergraduate composer.

4. AN ALGORITHMIC JURY TO ASSESS A REMIX COMPETITION

The formation of not just a single algorithmic critic, but a jury of critics, models many human competitive situations where outcomes are negotiated amongst multiple parties. It also evokes the dangers of decision by committee within a more complicated mechanics of bias; choosing the weightings between multiple agents is a critical aspect of ensemble methods in machine learning algorithms [Dietterich 2000; Tresp 2001].

The case study described here arose from a real world opportunity in judging a dubstep remix competition, and led to somewhat hyperbolic real world press coverage [Shaw 2012]. The source code for an algorithmically composed piece by Kiti le Step was made available for remixing, as well as its audio stems and the original track (http://www.sc2012.org.uk/2012/02/the-supercollider-algostep-remix-competition/). The conceit was that since the original work had been algorithmically composed, and remix competition entries were being created by algorithmic processes, why not judge the whole competition by machine?

Rather than a single judge, three algorithmic judges were convened; each judge had their respective biases, being trained on different audio databases. Table 2 lists the three judges and their respective training corpora; the artificial personalities are limited, and should not be overly anthropomorphized. Judges actually used three models each, one for their primary professionally stated likes as main training corpus, one anti-corpus of strongly disliked music (their 'pet hate'), and a secretive 'guilty pleasures' factor to introduce additional richness to their machine personality. In the final reckoning, judges held equal weight on the panel.

Judge	Main corpus	Pet hate	Guilty pleasure	
Judge Rules	General electronic dance music, 1990s popular dance music	The Beatles	Erasure	
Critex	Commercial dubstep and brostep	Schoenberg	Aphex Twin	
Code Fine	Earlier Croydon dubstep	Commercial dubstep and brostep	Abba	

Table 2: The algorithmic jury and their training sets

As in the previous section, predictive models were prepared using feature extraction, dimension reduction and discretization to an integer sequence, and modeling of those sequences via prediction by partial match (as in section 3, using order 5 models). However, beyond the technique of the previous section, three parallel models were prepared for each corpus, one each for timbre, pitch and rhythmic features (thus, 27 models were required for the jury). The features extracted and the discretization method utilized prior to each prediction by partial match model are listed in Table 3.

A further refinement was the addition of a compression measure, based on the lossless encoding scheme Lempel–Ziv–Welch (LZW compression; such compression algorithms have a close link to symbolic predictive modeling such as prediction by partial match [Begleiter at al. 2004]). LZW compression was used to downweight audio files which were too simplistic, that is, to exclude competition entries which were overly easy to predict. A neutral corpus of 30 historic synth pop tracks was used to train three parallel models again as per Table 3. For each competition entry, the integer sequence of class states $\{x_i\}$ for each of timbre, rhythm and pitch with respect to the synth pop model, were put through LZW compression to see to what degree the compression could reduce storage size Φ of the sequence:

$$complexity(\{x_i\}) = \frac{\Phi(LZW(\{x_i\}))}{\Phi\{x_i\}}$$

where complexity takes on values from 0.0 to 1.0, with maximal 1.0 corresponding to no possible compression (prediction) of a sequence.

Musical aspect	Features extracted	Discretisation
Timbre	perceptual loudness, transient information at two thresholds, Sethares sensory dissonance, spectral centroid, 80% and 95% spectral percentile, zero crossing rate, spectral crest factor, spectral slope, complex domain onset detection function (11 features)	k-Means, k=20
Pitch	harmonic change (summed framewise difference of 12 chromagram entries)	quantized to one of twenty histogram bins
Rhythm	inter-onset-intervals resulting from onset detection on the audio file	quantized to one of twenty histogram bins

Table 3: Feature space used for algorithmic jury models

Scores arose in each case from average log loss according to the model of a new input sequence, divided by the complexity score (values less than 1.0 thus increase dissimilarity). A model's score can be thought of as a particular similarity measure, ascertaining match of the new input audio file to a judge's "preferences" (with inverted scores to dissimilarity for the dislikes). The compression measure multiplier disadvantages pathological cases, e.g., audio with a lot of silence or simple periodic signals, but for the majority of "normal" pieces takes on similar values. Scores were normalized by max and min over all competition entries for a particular model to obtain values from 0.0 to 1.0, inverting as necessary (1- score) so that high similarity scored highly. Timbre, pitch and rhythm models were combined with equal weights, but the judge's main corpus had a much higher weighting (70%) than the pet hates (10%) and guilty pleasures (10%). The remaining 10% in each case was provided by an additional factor introduced to model the basic compromise of creativity on the Wundt curve [Deliège and Wiggins 2006], that is, that a remix should be close but not too close to the track being remixed. A further model was derived from the track to be remixed, to apportion its proximity to a given entry. After normalization of scores, each of the three judges took the linearly interpolated value between 1-score and score at proportions of 0.0, 0.5, and 0.75 respectively; that is, Judge Rules was extremely cautious (0.0) in “wanting” submissions closest to the original, Critex was moderate (0.5), and Code Fine looked to get the further from the original itself (0.75).

By the close of the competition, which had stated up front that an algorithmic jury would select the winner, 15 entries had been received. No human listened to the entries ahead of the program run. The initial predictive model formation ran overnight, training in about 6 hours. The algorithmic jury was in session considering the actual competition entries for just over an hour. The pragmatics of the hard deadlines of the competition close and announcement of winner limited the experimentation that could take place. Indeed, the actual competition entries could not be used for testing ahead of the final run, lest a bias be introduced favoring any one. Instead, some proxy electronica audio files were used to check the mechanisms of the jury run.

The winners of the competition were announced at an international conference, The SuperCollider Symposium 2012 in London held at Queen Mary University of London, with associated media coverage from the BBC’s Click program. The top three remixes were made available online and can be auditioned by the reader (http://chordpunch.com/2012/04/sc2012-kiti-le-step-competition-results/). The overall winner, all n4tural's *Kiti From Occupied Europe* solicited some surprise; it is a rather more experimental rhythmically and timbrally disjointed piece than other entries. It’s as if the computer selected ‘one of its own’, a work somewhat unconventional and inhuman in its timing. The selection made it clear, though, that much work remains to be done on the musical modeling underlying the process. A more developed critical process might involve a series of ‘sanity checks’ for stylistic appropriateness of submissions based on consistency of metrical structure and musical flow within whole pieces.

Once the computer decision was made, the jury designer and the creator of the original track to be remixed rated the entries themselves. Their results were predictably divergent, not only with the computer, but with each other. A contestant might reasonably ask, given the many decisions on features to be extracted, and weightings between models to determine final ranking scores, could the results have turned out any other way? The answer is definitely yes, within limits. Figure 2 shows the effect on the final score of varying the main corpus weighting within the calculation; whilst for very low main weight, there would have been a different winner, the top two positions quickly become clear, even if the third position is closer run (note that whilst lines appear straight, there are small fluctuations for each piece; the lowest scoring piece matched up poorly to all databases used). Participants had accepted the plan up front, and accepted the outcome of this algorithmic judging; a reasonable attempt had been made to simulate the jury process. At least the computational jury wasn’t biased by the time of day and how full their belly was, and there was no attempt to simulate the accumulation of listening fatigue over surveying many entries.

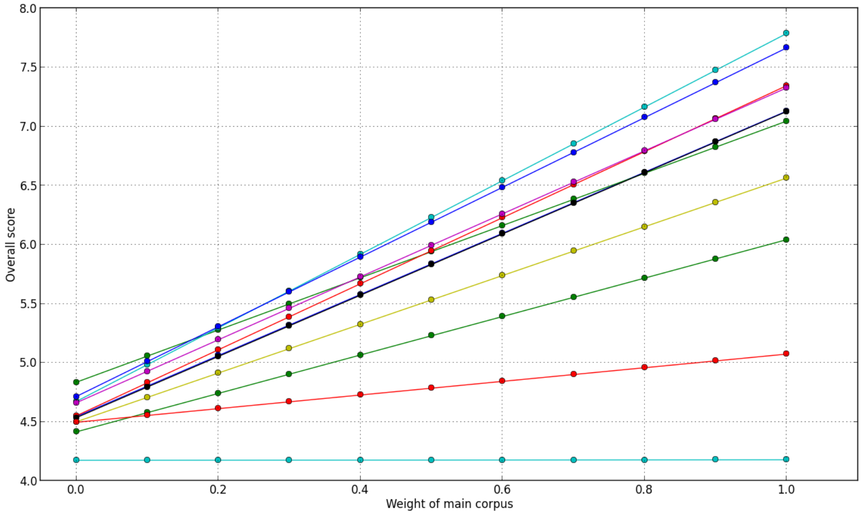

Figure 3 Variation in final overall score for the 15 competition entries (one per line) as the main corpus weighting is changed from 0.0 to 1.0 in steps of 0.1

5. LIVE APPLICATION

Concert criticism during a concert is increasingly available, though not widely adopted. Social media channels can provide in-concert feedback from human peers; network music pieces might incorporate a live twitter feed. The mood of an artificial musical instrument, FOUND collective's *Cybraphon*, is modulated by online attention [Found 2009]; a variant of Al Bile's GenJam can evolve based on audience feedback (Biles 2007). Performance evaluation with immediate or follow-on scoring is also a natural part of musical education applications, from video games to practice tutors [Dittmar 2012]. The metrics therein, however, are typically simpler, concerning constrained practice tasks with respect to known and accepted scoring mechanisms, rather than the challenge of computational creativity modeling.

It is perfectly feasible to run the algorithmic critic systems described in the preceding two sections live on audio input; the software utilized to construct them is native to a computer music environment inherently well suited to live performance, and the calculation demands on a single audio stream are not especially high (typically, around 10-20% of CPU, depending on the number and complexity of audio features extracted). Appropriate windowing decisions need to be made about what constitutes a judgment of music in the moment; average log loss calculations can operate on shorter integer sequences, and would work over a window of the previous N seconds, where N=10 might be the limits of short-term memory [London 2012].

As an example of the construction of an automated critic system specifically for live application, a live critic which measures proximity of incident musical material to a corpus of Keith Jarrett is described. The context is solo piano improvisation. The stages to building the system are as follows:

1. Collate a corpus consisting of the *Vienna Concert* (1991), and the second disc of *Radiance* (2002), tracks X to XVII, to form a two hour audio database
2. Extract features over the corpus. The 22 features utilized are a twelve pitch class chromagram, spectral entropy, spectral centroid, 75% and 25% spectral percentiles, spectral crest factors at 10kHz and 5kHz, predominant fundamental frequency detected from a (monophonic) pitch detector as well as detection confidence, perceptual loudness, key clarity (as a measure of how clearly pitch content matches to a single major or minor key)
3. Normalize features with respect to global minimum and maximum values
4. Cluster the feature vectors using a kMeans clusterer with k=26 (to create an alphabet of symbols)
5. Replace feature vector sequences from the corpus with kMeans cluster assignments, creating discrete symbolic sequences representing the music
6. Model the discrete symbol sequences so derived using a Markov model (order 5 prediction by partial match)
7. Use the model so formed as a predictive model. New audio input is converted into feature vectors, and thence to symbols via the same kMeans clusterer as above. Symbol sequences are tested for their predictability with respect to the existing Markov model

This method leads to a running value measuring the degree of predictability of novel input to the system, with respect to the corpus of Keith Jarrett. Live, this value provides an interesting tension to the performer, who must negotiate their relationship with the master improviser Jarrett. The single numerical value can be normalized to the range 0-1, based on expected ranges (given previous recordings of the performer), or adaptively online in-concert based on the maximum and minimum observed so far. The value can also be associated with a corpus of positive or negative words, printed live, with the mapping determined by whether proximity to Jarrett is seen as a positive or a negative in itself.

As illustration of the model output, Figure 3 plots the average log loss (without final normalization), for a Keith Jarrett piece (*Radiance Part XV*, black dashed line) versus a third party novel piano improvisation (green line), also of around ten minutes. The performer was unaware of the Jarrett model being applied (the recording having been made some years prior to this research project), though their base style of play worked within a contemporary improvisation paradigm much influenced by Jarrett. The Jarrett improvisation is better predicted by the model as shown by the lower average log loss values; further, those areas of the novel improvisation which are better predicted (higher 'Jarrettosity', lower y value), on a qualitative listen, show a musical connection. Particularly around 480 seconds in, the novel improvisation enters a more mellow homophonic chord progression; earlier regions of low y value (such as 5, 115, 280 seconds in) show reduced complexity of playing; the more predominant material in the novel piano performance involves extended techniques such as playing inside the piano, and other contemporary classical improvisation further from the sorts of jazz and tonal improvisatory practice associated with Jarrett. Yet, no claim is possible here to have captured high-level critical aspects of Jarrett's playing, and signal confounds such as low complexity moments of silence (leading to higher predictability) have an effect.

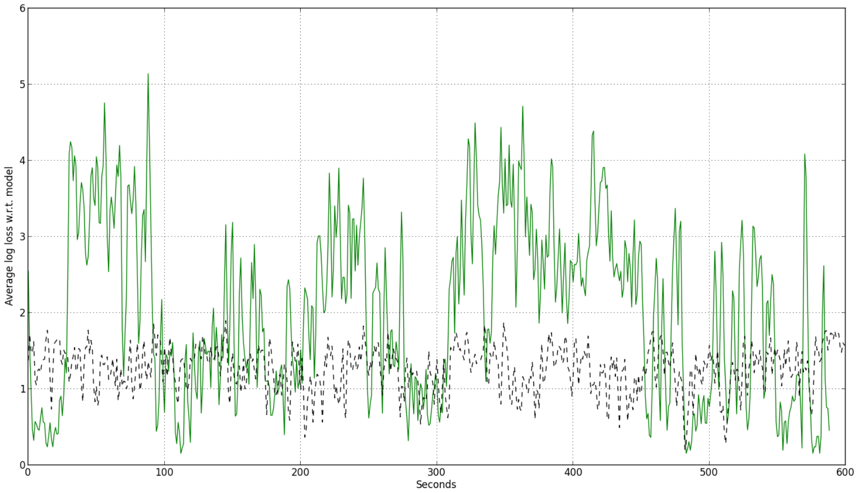

Figure 4 Keith Jarrett audio trained model evaluating a Keith Jarrett improvisation (black dashed line) and a third party performance (green continuous line), both of around ten minutes. The lower the value, the better predicted the music and the greater the putative"Jarrettosity"

6. DISCUSSION

In all the experiments detailed in this article, an interaction of subjective and objective is inescapable. Representational decisions and material used for training establish bias, introduced by the system designer whatever their ostensible desire to formalize musical judgment. This is true of human critics, too, based on their own favored musical preferences; in a recent doctoral thesis, Christopher Robinson [2014] notes the ‘multiple insularities’ of jazz criticism, with individual jazz critics selecting different canonic repertoire, chosen according to personal agendas. Forming a corpus representative of a particular style is a heady challenge, and can necessitate cross-referenced study of multiple sources [London 2013], one motivation in the earlier stage of this paper to simultaneously utilize two corpuses.

Research on artificial critics trained on larger audio databases is at a tentative time. But whenever machine critics are more widely adopted, beyond the current state of the art, there will be interesting consequences in guiding human cultural evolution mediated by the reification of particular machine models. The level of artificial bias to add to any model is a dangerous choice; modeling of humanity might require enough obvious bias to provide character (as was anthropomorphized in section 4), yet perhaps the full bias of an artificial agent needs to be hidden from itself and from other agents to act in a human-like, less than fully self-aware way within culture.

The critics created in this project so far have not been able to communicate their judgments in a more involved human mode, such as programme notes, or a written review. As noted in section 2, existing projects have attempted to connect actual review texts to audio feature data through correlation of occurrence [Whitman and Ellis 2004]; it is certainly plausible to build such a system for real-time performance use. Nonetheless, a deeper musical analysis and criticism, in the sense of more human-like machine listening, and the deeper conception of critical analysis of music touched on at the beginning of section 2, remains a strong research challenge. For future work, the notion of a human critic's computational critic assistant spotting precursors and other musical relations, or a human creator's automated composition assistant providing reliable feedback on the pure level of innovation, are more manageable goals, though still open to skepticism of the level of listening encapsulated within their design.

It is always possible to point to doubts on current generation machine listening capability. There are some more advanced features that might have been applied. For the remix competition judging, it would have been feasible to extract chord sequences, and even run a polyphonic pitch tracker; nonetheless, error rates particularly of the latter remain, and there would still be a sense of machines lacking a human-like high-level auditory object discrimination / musical stream segmentation capability. Critical judgments are not returned and justified with respect to high level musical objects, components of musical structure noted by human theorists, but instead lower level features, discretization and symbolic sequence maps interpose. On the other hand, since a human critic might earn trust over time by convincing a particular audience of the clarity of their assertions, so we may come to trust particular algorithmic critics, even be pleased by their quirks. Human critics have their own vagaries; in a 1915 scandal, the Russian critic Sabaneyev reviewed Prokofiev's *Scythian Suite*, unaware of the premiere's cancellation since he never had any intention of turning up to hear it in the first place [Prokofiev and Phillips 2008]!

It is well known that the pursuit of novelty for the sake of novelty is not the sweet spot of a Wundt curve of creativity [Deliège and Wiggins 2006; McCormack and d'Inverno 2012], and so a balanced relationship to previous models is necessary for accessible yet exciting creation. With the algorithmic jury of section 4, the three virtual critics had their own 'opinion' of the best target mix of matching a piece and moving to the fringes. Indeed, without a much wider modeling of the operational space of music (as alluded to with the section 3 study), the notion of what might constitute 'near but not too near' is less well defined.

Having created three distinct systems, we return to Spector and Alpern's three challenges [Spector and Alpern 1994]. This article has dealt with music which is alive rather than dead: continuing developments in popular and art electronic music, and recent piano improvisation. All work with similarity measures ultimately led to a single dimension score, and we have outlined the recurrent problems of the scope of aesthetic inclusion and exclusion. Where is the dividing line for safe membership, and where is it for edgy work? By working with larger audio databases, particularly in the double electronic music corpus study of section 3, greater robustness is sought, though there remain the issues of acceptable criteria. We might hypothesize that the problem of determining such criteria will always remain in any cultural evolution which must keep in touch with human activity. There is no ultimate aesthetic solution, but we can still do better at the musical modeling underlying this work.

7. CONCLUSIONS

In this article, newly generated electroacoustic pieces were measured up to the curated and established work of human electroacoustic composers associated with the respected empreintes DIGITALes record label, or with a corpus of historical electronic art music. An objective, though no less biased algorithmic jury was formed to judge a real world remix competition. A live system was discussed which forms a stimulation, or obstruction, depending on your attitude, to live piano improvisation. All of these projects are enabled by MIR techniques on larger audio databases; concerns on the level of machine listening remain a basic research challenge, but it is salutary to see where the technology might take us.

Some musicians may worry about the danger of guidance from trusting an artificial musical proxy; others may embrace the opportunity. Media exposure, current and future, can only promote metacreative systems within musical culture. Research in this area is of interest as a further viewpoint on the socialization of musical machines, providing a challenging test case for artificially intelligent musicianship.

ACKNOWLEDGMENTS

With thanks to Dan Stowell in particular for his involvement in running and publicizing the algorithmically juried competition. Acquisition of the empreintes DIGITALes corpus was funded from a grant from Durham University; Jean-François Denis assisted with providing the back catalogue in digital form.

REFERENCES

Begleiter, R., El-Yaniv, R., and Yona, G. (2004). On prediction using variable order Markov models. *Journal of Artificial Intelligence Research* 22: 385-421

Bertin-Mahieux, T., Eck, D., and Mandel, M. (2010) Automatic tagging of audio: The state-of-the-art. *Machine Audition: Principles, Algorithms and Systems*. IGI Publishing

Biles, J. A. (2007) Improvising with genetic algorithms: GenJam. pp. 137-169 in Miranda, E., and Biles, A. (eds.) *Evolutionary Computer Music*. London: Springer

Casey, M., and Slaney, M. (2006) The importance of sequences in musical similarity. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C. and Slaney, M. (2008) Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE* 96(4): 668–96

Collins, N. (2011) Musicians' Attitudes to Musical Influence. *Empirical Musicology Review* 6(2): 103-124

Collins, N. (2012) Automatic Composition of Electroacoustic Art Music Utilizing Machine Listening. *Computer Music Journal* 36(3): 8-23

Collins, N. (2015) The UbuWeb Electronic Music Corpus: An MIR investigation of a historical database. *Organised Sound* 20(1): 122-134

Compton, K., Osborn, J. C., and Mateas, M. (2013) Generative methods. In *The Fourth Procedural Content Generation in Games workshop, PCG*.

Cope, D. (ed.) (2001). *Virtual Music: Computer Synthesis of Musical Style*. MIT Press, Cambridge, MA.

Deliège, I. and Wiggins, G. A. (eds.) (2006) *Musical Creativity: Multidisciplinary Research in Theory and Practice*. Hove: Psychology Press

Dietterich, T. G. (2000) Ensemble Methods in Machine Learning. *Lecture Notes in Computer Science* 1857: 1-15

Dittmar, C., Cano, E., Abeßer, J., and Grollmisch, S. (2012) Music Information Retrieval Meets Music Education. pp. 95–120 in M. Muller, M. Goto, and M. Schedl (eds.) *Multimodal Music Processing*, volume 3 of Dagstuhl Follow-Ups, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany

Fernández, J. D., and Vico, F. (2013) AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research* 48: 513-582.

Found (2009) Cybraphon. *National Museums Scotland*. Online at http://www.nms.ac.uk/explore/collections-stories/science-and-technology/cybraphon/

Galanter, P. (2012) Computational Aesthetic Evaluation: Past and Future. In Jon McCormack and Mark d'Inverno (Eds.) *Computers and Creativity*. Berlin: Springer

Hu, X., Downie, J. S., West, K., and Ehmann, A. F. (2005) Mining Music Reviews: Promising Preliminary Results. In *Proceedings of the International Symposium on Music Information Retrieval* (ISMIR), London

Jewell, M. O., Rhodes, C., and d'Inverno, M. (2010) Querying improvised music: Do you sound like yourself? In *Proceedings of the International Symposium on Music Information Retrieval* (ISMIR), Utrecht

London, J. (2012) *Hearing in time*. New York: Oxford University Press

London, J. (2013) Building a representative corpus of classical music. *Music Perception* 31(1): 68-90.

Machado, P., Romero, J., Manaris, B., Cardoso, A., and Santos, A. (2003) Power to the Critics – A Framework for the Development of Artificial Art Critics. In *Proceedings of 3rd Workshop on Creative Systems, 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, Acapulco, Mexico, Aug. 2003.

McCormack, J., and d'Inverno, M. (eds.) (2012) *Computers and Creativity*. Berlin: Springer

Meyer, L. B. (1973) *Explaining music: Essays and explorations*. Univ of California Press.

Pearce, M. and Wiggins, G. (2004) Improved Methods for Statistical Modelling of Monophonic Music. *Journal of New Music Research* 33(4): 367–85

Prokofiev, S., and Phillips, A. (2008) *Sergey Prokofiev diaries, 1915-1923: behind the mask*. Ithica, NY: Cornell University Press

Ramsay, S. (2011) *Reading machines: Toward an algorithmic criticism*. Champaign, IL: University of Illinois Press

Robinson, C. (2014). *Firing the Canon: Multiple Insularities in Jazz Criticism*. Doctoral thesis, University of Kansas

Rowe, R. (1993) *Interactive Music Systems*. Cambridge, MA: MIT Press

Schick, R. D. (2013) *Classical Music Criticism*. New York: Routledge.

Spector, L., and Alpern, A. (1994) Criticism, culture, and the automatic generation of artworks. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pp. 3-8

Shaw, D. (2012) Cowell v computer: Have talent show judges had their day? *BBC News* 19 April 2012. http://www.bbc.co.uk/news/technology-17735551

Slaney, M., and Casey, M. (2008) Locality-sensitive hashing for finding nearest neighbors [lecture notes]. IEEE Signal Processing Magazine 25(2): 128-131

Stiny, G., and Gips, J. (1978) *Algorithmic aesthetics: computer models for criticism and design in the arts*. University of California Press

Sturm, B. L. (2014) The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. *Journal of new music research* 43(2): 147-172

Todd, P. M., and Werner, G. M. (1999) Frankensteinian methods for evolutionary music. pp. 313-340 in Griffith, N., and Todd, P. M. (eds.) *Musical networks: Parallel distributed perception and performance*. Cambridge, MA: MIT Press

Tresp, V. (2001) Committee Machines. In Yu Hen Hu and Jenq-Neng Hwang (eds.) *Handbook for Neural Network Signal Processing*. Boca Raton, FL: CRC Press

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (2008) Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing* 16(2):467–476

Virtanen, T., and Heleń, M. (2007) Probabilistic Model Based Similarity Measures for Audio Query-By-Example. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York

Whitman, B., and Ellis, D. P. (2004) Automatic record reviews. In *Proceedings of the International Symposium on Music Information Retrieval* (ISMIR 2004), Barcelona, Spain.

Received 1 May 2015, final revised version 8 March 2016, accepted late summer 2016