TOWARDS SEAMLESS NETWORK MUSIC PERFORMANCE: PREDICTING AN ENSEMBLE'S EXPRESSIVE DECISIONS FOR DISTRIBUTED PERFORMANCE

Bogdan Vera
Queen Mary University of London
Centre for Digital Music
b.vera@qmul.ac.uk

Elaine Chew
Queen Mary University of London
Centre for Digital Music
elaine.chew@qmul.ac.uk

ABSTRACT

Internet performance faces the challenge of network latency. One proposed solution is music prediction, wherein musical events are predicted in advance and transmitted to distributed musicians ahead of the network delay. We present a context-aware music prediction system focusing on expressive timing: a Bayesian network that incorporates stylistic model selection and linear conditional gaussian distributions on variables representing proportional tempo change. The system can be trained using rehearsals of distributed or co-located ensembles.

We evaluate the model by comparing its prediction accuracy to two others: one employing only linear conditional dependencies between expressive timing nodes but no stylistic clustering, and one using only independent distributions for timing changes. The three models are tested on performances of a custom-composed piece that is played ten times, each in one of two styles. The results are promising, with the proposed system outperforming the other two. In predictable parts of the performance, the system with conditional dependencies and stylistic clustering achieves errors of 15ms; in more difficult sections, the errors rise to 100ms; and, in unpredictable sections, the error is too great for seamless timing emulation. Finally, we discuss avenues for further research and propose the use of predictive timing cues using our system.

1. INTRODUCTION

Ensemble performance between remote musicians playing over the Internet is generally made difficult or impossible by high latencies in data transmission [3] [5]. While many composers and musicians have chosen to treat latency as a feature of network music, performance of conventional music, such as that of classical repertoire, remains extremely difficult in network scenarios. Audio latency frequently results in progressively decreasing tempo and difficulty in synchronizing.

One aspect that has received less attention than the latency is the lack of visual contact when performing over the internet. Visual cues can be transmitted via video, but such data is at least as slow as audio, and was previously found to not be of significant use for transmitting synchronization cues even when the audio had an acceptable latency [6].

Since the start of network music research, several researchers have posited theoretically that music prediction could be the solution to network latency (see, for example, Chafe [2]). Ideally, if the music can be predicted ahead of time with sufficient accuracy, then it can be replicated at all connected end-points with no apparent latency. Recent efforts have made limited progress towards this goal. One example is a system for predicting tabla drumming patterns [12], and recent proposals by Alexandraki [1]. Both assume that the tempo of the piece will be at least locally smooth and, in the case Alexandraki's system, timing alterations are always based on one reference recording.

In many styles of music, such as romantic classical music, the tempo can vary widely, with musicians interacting on fine-scale note-to-note timing changes and using visual cues to synchronize. The tempo cannot be expected to always evolve in the exact same way as one previous performance, rather the musicians significantly improvise timing deviations to some constraints.

In this paper we propose a system for predicting timing in network performance in real time, loosely inspired by Raphael's approach based on Bayesian networks [11]. We propose and test a way to incorporate abstract notions of expressive context within a probabilistic framework, making use of time series clustering. Flossman et al. [8] employed similar ideas when they extended the YQX model for expressive offline rendering of music by using conditional gaussian distributions to link expressive predictions over time. Our model contains an extra layer of stylistic abstraction and is applied to modeling and real-time tracking of one performer or ensemble's expressive choice at the inter-onset interval level. We also describe how the method could be used for predicting musical timing in network performance, and discuss ideas for further work.

2. MOTIVATION

Our goal is to use observable sources of information during a live performance to predict the timing of future notes so as to counter the effects of network latency. The sources of information we can use include the timing of previous notes and the intensity with which the notes are played.

The core idea is reminiscent of Raphael's approach to automatic accompaniment [11], which uses a Bayesian network relating note onset times, tempo and its change over time. In Raphael's model, changes in tempo and local note timing are represented as independent gaussian variables, with distributions estimated from rehearsals. During a performance, the system generates an accompaniment that emulates the rehearsals by applying similar alterations of timing and tempo at each note event in the performance. The model has been demonstrated in live performances and proven to be successful, however as long as the system generates musically plausible expression in the accompaniment, it is difficult to determine an error value, as it is simply meant to follow a musician and replicate a performance style established in rehearsals. An underlying assumption of this statistical model is that the solo musician leading the performance tends to perform the piece with the same expressive style each time.

In an ensemble performance scenario, two-way communication exists between musicians. The requirement for the system to simply 'follow' is no longer enough. As a step towards tighter ensemble, we set as a goal a stringent accuracy requirement for our prediction system: to have errors small enough—no higher than 20-40ms—as to be indistinguishable from the normal fluctuations in ensemble playing. Note that actual playing may have higher errors, even in ideal conditions, due to occasional mistakes and fluctuations in motor control.

The same ensemble might also explore a variety of ways to perform a piece expressively. When expressive possibilities are explored during rehearsals, the practices establish a common 'vocabulary' for possible variations in timing that the musicians can then anticipate. Another goal of our system is to account for several distinct ways of applying expression to the same piece. This is accomplished in two ways. Like Flossman et al. [8], we deliberately encode the context of the local expression by introducing dependencies between the expressive tempo changes at each time step. We additionally propose and test a form of model selection using discrete variables that represent the chosen stylistic *mode* of the expression. For example, given two samples exhibiting the same tempo change, one may be part of a longer term tempo increase, while another may be part of an elastic time-stretching gesture. Knowing the stylistic context for a tempo change will allow us to better predict its trajectory.

3. CONTEXTUALIZING TIMING PREDICTION

We combine two techniques to implement ensemble performance prediction. First, we condition the expressive 'update' distributions characterizing temporal expression on those from preceding events, making the timing changes dependent on both musicians' previous timing choices, while also allowing the system to respond to the interplay between the two musicians. Secondly, we abstract different ways of performing the piece by summarizing these larger scale differences in an unsupervised manner in a new discrete node in the network: a stylistic cluster node.

3.1 Linear Gaussian Conditional Timing Prediction

Our goal is to predict the timing of events such as notes, chords, articulations, and rests. In particular, we wish to determine the time until the next event given the score information and a timing model. We collapse all chords into single events. Assume that the performance evolves according to the following equations,

$$\begin{aligned} t_{n+1} &= s_n l_n + t_n, \text{ and} \\ s_{n+1} &= s_n \cdot \delta_n, \end{aligned} \tag{1}$$

where t_n is the onset time of the n-th event, s_n is the corresponding inter-beat period, l_n is the length of the event in beats, and δ_n is a proportional change in beat duration that is drawn from the gaussian distributions Δ_n. For simplicity, there is no distinction between tempo and local timing in our model, though it could be extended to include this separation.

Because δ_n's reflect proportional change in beat duration, prediction of future beat durations are done on a logarithmic scale:

$$\log_2 s_{n+1} = \log_2 s_n + \log_2 \delta_n.$$

$\log(tempo) = \log(1/s_n)$, thus $\log s_n$ as well, has been shown in recent research to be a more consistent measure of tempo variation in expressive performance [4].

The parameters of the Δ_n distributions are predicted during the performance from previous observations, such as δ_{n-1}. Thus, each inter-beat interval, s_n, is shaped from event to event by the random changes, δ_n. The conditional dependencies between the random variables are illustrated in Figure 1. The first and last layers in the network, labeled P1 and P2 in the diagram, are the observed onset times. The 3rd layer, labeled 'Composite' following Raphael's terminology, embodies the time and tempo information at each event, regardless of which ensemble musician is playing, and it is on this layer that our model focuses. The 2nd layer, Expression, consists of the variables Δ_n.

The Δ_n variables are conditioned upon their predecessors, using any number of previous timing changes as input; formally, they are represented by linear conditional gaussian distributions [9]. Let there be a Bayesian network node with a normal distribution Y. We can condition Y on its k continuous parents $C = \{C_1, \ldots, C_k\}$ and discrete parents $D = \{D_1, \ldots, D_k\}$ by using a linear regression model to predict the mean and variance of Y given the values of C and D. The following equation describes the conditional probability of Y given only continuous parent nodes:

$$P(Y|C = \mathbf{c}) = \mathcal{N}(\beta_0 + \sum_{i=1}^{k} \beta_i c_i, \sigma^2).$$

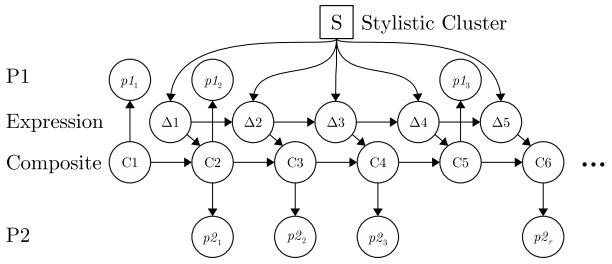


Figure 1: A section of the graphical model. Round nodes are continuous gaussian variables, and the square node (S) is a discrete stylistic cluster node.

This is the equation for both continuous and discrete parents:

$$P(Y|D=\mathbf{d},C=\mathbf{c}) = \mathcal{N}(\beta_{\mathbf{d},0} + \sum_{j=1}^{k} \beta_{\mathbf{d},j} c_j, \sigma^2_{\mathbf{d}}).$$

Simply speaking, the mean and variance of each linear conditional gaussian node is calculated from the values of its continuous and discrete parent nodes. The mean is derived through linear regression from its continuous parents' values with one weight matrix per configuration of its discrete parents.

The use of conditional gaussian distributions means that rather than having fixed statistics for how the timing should occur at each point, the parameters for the timing distributions are predicted in real time from previous observations using linear regression. This simple linear relationship provides a means of predicting the extent of temporal expression as an ongoing gesture. For example, if the performance is slowing down, the model can capture the rate of slowdown, or a sharp tempo turnaround if this occurred during rehearsals.

Our network music approach involves interaction between two actual musicians rather than a musician and a computer. Thus, each event observed is a 'real' event, and we update the Δ_n probability distributions at each step during run-time with the present actions of the musicians themselves. Unlike a system playing in automatic accompaniment or an expressive rendering system, our system is never left to play on its own, and its task is simply to continue from the musicians' choices, leaving less opportunity for errors to accumulate. Additionally, we can correct the musicians' intended timing by compensating for latency post-hoc - this implies that we can make predictions that emulate what the musicians would have done without the interference of the latency.

We may also choose the number of previous changes to consider. Experience shows that adding up to 3 previous inputs improves the performance moderately, but the performance decreases thereafter with more inputs. For simplicity, we currently use only one previous input, which provides the most significant step improvement.

In constrast to a similar approach by Flossman et al. [8], we do not attempt to link score features to the performance; we only consider the local context of their temporal expression. Our goal is to capture the essence of one particular ensemble's interpretation of a particular piece rather than attempting to construct a universal model for mapping score to performance. As a result, the amount of training data will generally be much smaller as we may only use the most recent recorded and annotated rehearsals of the ensemble. The next section describes a clustering method we use to account for large-scale differences in timing.

3.2 Unsupervised Stylistic Characterization

Although we could add a large number of previous inputs to each of the Δ_n nodes, we cannot tractably condition these variables' distributions on potentially hundreds of previous observations. This would require a large amount of training data to estimate the parameters in a meaningful way. Instead, we propose to summarize larger-scale expression using a small number of discrete nodes representing the stylistic mode. For example, a musician may play the same section of music in few distinct ways, and a listener may describe it as 'static', 'swingy' or 'loose'. If these playing styles could be classified in real time, prediction could be improved by considering this stylistic context. Our ultimate goal is to perform this segmentally on a piece of music, discovering distinct stylistic choices that occured in the ensemble's rehearsals. In this paper, we present the first steps towards this goal: we characterize the style of the entire performance using a single discrete stylistic node.

The stylistic node is shown at the top of Figure 1. In our model this node links to all of the Δ_n nodes in the piece, so that each of the Δ_n's is now linearly dependent on the previous timing changes with weights that are dependent on the stylistic node. Assuming that each Δ_n node is linked to one previous one, the parameters of the Δ_n distributions are then predicted at run-time using

$$P(\Delta_t|S=s,\Delta_{t-1}=\delta) = \mathcal{N}(\beta_{s,0} + \beta_{s,1}\delta, \sigma^2_s),$$

where S is the style node.

To predict note events, we can simply take the means of the Δ_n distributions, and use Equation 1 to find the onset time of the next event given the current one.

To use this model, we must first discover the distinct ways (if any) in which the rehearsing musicians perform the piece. We apply k-means clustering to the $log(\delta_n)$ time series obtained from each rehearsal. We find the optimal number of clusters by using the Bayes Information Criterion (BIC) as described by Pelleg and Moore [10]. Note that other methods exist for estimating an optimal number of clusters. To train the Bayesian network, a training set is generated containing all of the δ_n values for each rehearsal as well as the cluster to which each time series is allocated. We then use the algorithm by Murphy [9] to find all the parameters of the linear conditional nodes. Note that all of the nodes are observable and we have training data for the Δ_n.

During the performance, the system can update its belief about the stylistic node's value from the note timings that have been observed at any point; we do not need to re-cluster the performance, as the network has learned the relationships between the Δ_n's and the stylistic node. We

use the message passing algorithm of Bayesian networks to infer the most likely state of the node. As the performance progresses, the belief about the state of the node is gradually established. Intuitively, the system arrives at a stable answer after some observations, otherwise the overall style is ambiguous. The state of the node is then used to place future predictions into some higher level context. The next section shows that the prediction performance is improved by using the stylistic node to select the best regression parameters to predict the subsequent timing changes, which can be thought of as a form of model selection.

4. EVALUATION

4.1 Methodology

In this section we present an evaluation of the basic form of our model. Evaluation of such predictive models remains a challenge because testing in live performance requires further work on performance tracking and optimization, while offline testing necessitates a large number of annotated performances from the same ensemble. We present initial results on a small dataset; in our future work we will study real time performances of more complex pieces.

We evaluate the performance of three models: one uses linear conditional nodes and a stylistic cluster node; the second uses only linear conditional nodes; and, the third has independent gaussian distributions for the Δ variables.

Our dataset consists of 20 performances by one pianist of the short custom-composed piece shown in Figure 2. Notice that we have not added any dynamics or tempo-related markings - the interpretation is left entirely to the musicians. While this is not an ensemble piece, the performances are sufficient to test the prediction accuracy of our model in various conditions. In this simple example, we consider only the composite layer in the model, without P1 and P2.

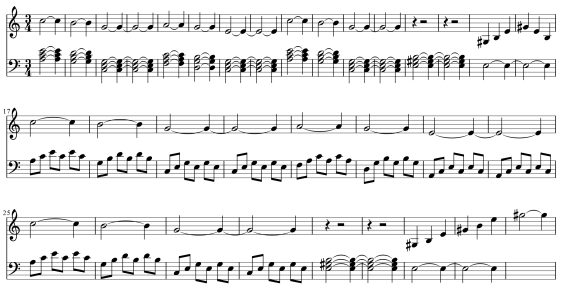


Figure 2: Custom-composed piano test piece.

The piece was played on an M-Audio AXIOM MIDI keyboard in one of two expressive styles decided beforehand, ten times for each style. We used IRCAM's Antescofo score follower [7] for live tracking of the performance in our system, and annotation of the note and chord events. The log-period plots for every performance in the dataset are shown in Figure 4a. The changes in log-period per event are shown in Figure 4b, and we also show the same changes but for the data in each cluster found, to demonstrate the difference between the two playing styles.

We evaluated the system using a 'leave-one-out' approach, where out of the 20 performances we always trained on 19 of them and tested on the remaining one. We always used one previous input to the Δ_n nodes, using the actual observations in the performances rather than our predictions (like the extended YQX), simulating the process of live performance. We evaluated the prediction accuracy by measuring timing errors, which we define as the absolute difference between the true event times and those predicted by the model (in seconds).

The training performances were clustered correctly in all cases, dividing the dataset into the two styles, with the first 10 performances being grouped with cluster 1 and the second 10 becoming part of cluster 2. Figure 3 shows the stylistic inference process. In the matrix, performances are arranged as rows, with events on the x-axis. Recall that we predict the time between events rather than just notes. So, we also consider the timing of rests, and chords are combined into single events rather than individual notes. The colors indicate the inferred value of the style node: grey for Style 1 and white for Style 2. We see that the system correctly infers the stylistic cluster of each performance within the first 19 events. In many cases the classification assigns the performance to the correct cluster after only two events.

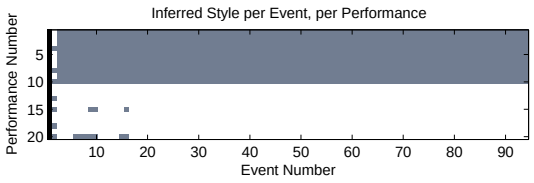


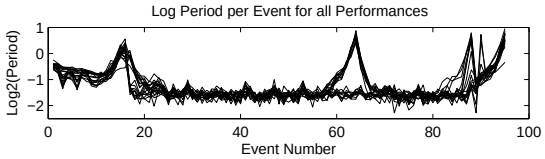
Figure 3: Matrix showing most likely style state after each event's observed δ. Performances 1-10 are in Style 1, and 11-20 are in Style 2. Classification result: grey = Style 1, white = Style 2.

Figure 4 shows the tempo information for the dataset. Figure 4(a) shows the inter-beat period contours of all of the performances, while Figure 4(b) shows boxplots (indicating the mean and variability) of the period at each musical event, for the entire dataset and for the two clusters.

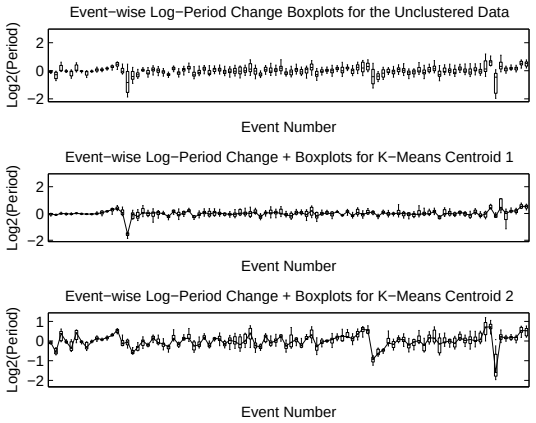
4.2 Results

Figure 5a and Figure 5b show the performance of the models, measured using mean absolute error averaged over events in each performance, and over performances for each event, respectively. We also show a detailed 'zoomed in' plot of the errors between events 20-84 to make the different models' mean errors clearer in Figure 5c. For network music performance, we would want to predict at least as far forward as needed to counter the network (and other system) latency. As some inter-event time differences may be shorter than the latency, we may occasionally need to predict more than one event ahead.

The model with stylistic clustering and linear conditional nodes performed best, followed by the one with only linear conditional nodes, then the model with independent

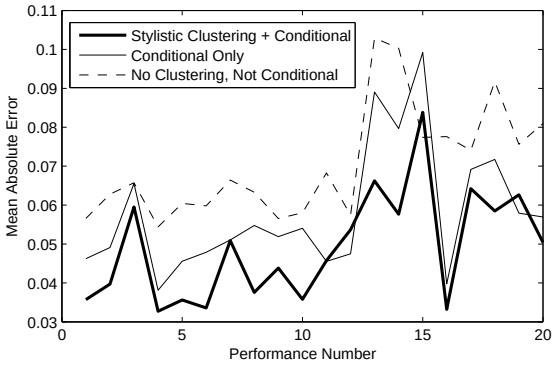


(a) Log-period per event for every performance in the dataset.

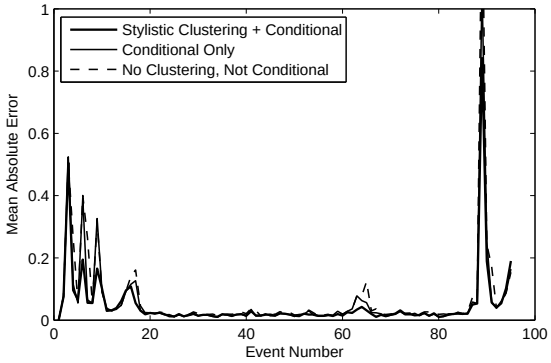


(b) Boxplots showing median and variability for the log-period change at each event. Top: unclustered data, Middle: first centroid, Bottom: second centroid.

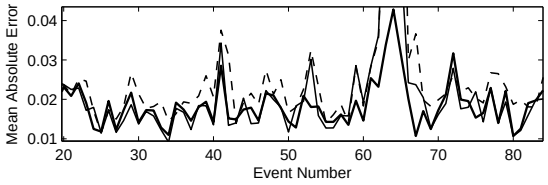
Figure 4: Tempo Data



(a) Mean absolute error for each performance.



(b) Mean absolute error per event, over the whole performance.



(c) A 'zoomed-in' view of the error rates between events 20-84.

Figure 5: Mean absolute error per event.

Δ_n nodes. In all cases the errors were higher for the second style (the latter 10 performances), which was much looser than the first. The mean absolute errors for each model, considering all of the events in all of the performances are summarized in Table 1.

Observe in Figure 5b that some parts of the performance were very difficult to predict. For example, we note high prediction errors in the first 12 events of the piece and one large spike in the error at the end of the piece. These are 1-bar and 2-bar long chords, for which musicians in an ensemble would have to use visual gestures or other information to synchronize. We would not expect any prediction system to do better than a musician anticipating the same timing without any form of extra-musical information. We discuss potential applications of music prediction for virtual cueing in the next section. The use of clustering and conditional timing distributions reduced the error rate for the events which were poorly predicted with independent timing distributions. For much of the piece the mean error was as low as 15ms, but even for these predictable parts of the performance, the models with conditional distributions and clustering lowered the error, as can be seen from Figure 5c.

5. CONCLUSIONS AND FUTURE WORK

We have outlined a novel approach to network music prediction using a Bayesian network incorporating contextual inference and linear gaussian conditional distributions. In an evaluation comparing the model with stylistic clustering and linear conditional nodes, one with only linear conditional nodes without clustering, and one with independent nodes, we have shown that the proposed approach produces promising results. Specifically, we have shown evidence that considering a notion of large scale expressive context, drawn from performance styles of a particular ensemble, can intuitively increase the accuracy of timing prediction. The model remains to be tested on more data. As creative musicians are infinitely diverse in their expressive interpretations, the true test of the model would ultimately be in live performances.

The end goal of this research is to implement and evaluate network music performance systems based on the prediction model. Whether music prediction can ever be precise enough to allow seamless network performance remains an open question. Important questions arise in pur-

Model	Mean Abs. Error
Independent	69.8ms
Conditional	57.4ms
Clustering and Conditional	48.5ms

Table 1: Overall Timing Errors for Each Model

suit of this goal: how much should the system lead the musicians to help them stay in time without making the performance artificial? Predicting musical timing with sufficient accuracy will open up interesting avenues for network music research, especially when we consider parallel research into predicting other information such as intensity and even pitch information, but whether any musician would truly want to let a machine impersonate them expressively remains to be seen, which is why we propose that a 'minimally-invasive' conductor-like approach to regulating tempo would be more appropriate than complete audio prediction.

5.1 The Bayesian Network

It would be straightforward to extend our model by implementing prediction of timing from other forms of expression that tend to correlate with tempo. For example, using event loudness in the prediction would simply require the addition of another layer of variables in the Bayesian network and conditioning the timing variables on these nodes as well.

5.2 Capturing Style

Much work remains to expand on the characterization of stylistic mode. As previously mentioned, we plan to explore segmental stylistic characterization, considering different contextual information for each part of the performance. In our current model we use only one stylistic node. This may be a plausible for a small segment of music, but in a longer performance the choice of performance style may vary over time. If the predicted performance starts within one style but changes to another, the model is ill-informed to predict the parameters. In our future work we would like to extend the model to capture such stylistic tendencies over time. One approach would require presegmentation of the piece based on the choice of expressive choices during the reharsal stage, and introduction of one stylistic node per segment. The prediction context would then be local to each part of the performance. We may then, for example, have causal conditional dependencies between the stylistic nodes in each segment of the piece, which would allow the system to both infer the style within a part of the performance from what is being played and from the previous stylistic choices.

In practice, a musician or ensemble's rehearsals may not comprise of completely distinct interpretations; however, capturing expression contextually will likely offer a larger degree of freedom to the musicians in an internet performance, who may then explore a greater variety of temporal and other articulations.

5.3 Virtual Cueing

Virtual cueing forms an additional application of interest. As mentioned at the start of the paper, visual communication is generally absent or otherwise delayed in network music performance. If we could predict with reasonable accuracy the timing in sections of a piece requiring temporal coordination, then we could help musicians synchronize by providing them with perfectly simultaneous predicted cues. We regard the use of predictive virtual cues as less invasive to networked ensembles than complete predictive sonification. In situations where the audio latency is low enough for performance to be feasible but video latency is still too high for effective transmission of gestural cues, predictive sonification may be omitted completely, and virtual cues could be implemented as a regulating factor.

6. ACKNOWLEDGEMENTS

This research was funded in part by the Engineering and Physical Sciences Research Council.

7. REFERENCES

[1] C. Alexandraki and R. Bader. Using computer accompaniment to assist networked music performance. In *Proc. of the AES 53rd Conference on Semantic Audio, London, UK*, 2013.

[2] C. Chafe. Tapping into the internet as an acoustical/musical medium. *Contemporary Music Review*, 28, Issue 4:413–420, 2010.

[3] C. Chafe and M. Gurevich. Network time delay and ensemble accuracy: Effects of latency, asymmetry. In *Proc. of the 117th Audio Engineering Society Convention*, 2004.

[4] E. Chew and C. Callender. Conceptual and experiential representations of tempo: Effects on expressive performance comparisons. In *Proc. of the 4th International Conference on Mathematics and Compution in Music*, pages 76–87, 2013.

[5] E. Chew, A. Sawchuk, C. Tanoue, and R. Zimmermann. Segmental tempo analysis of performances in user-centered experiments in the distributed immersive performance project. In *Proc. of the Sound and Music Computing Conference*, 2005.

[6] E. Chew, R. Zimmermann, A. Sawchuk, C. Kyriakakis, and C. Papadopolous. Musical interaction at a distance: Distributed immersive performance. In *Proc. of the 4th Open Workshop of MUSICNETWORK, Barcelona*, 2004.

[7] A. Cont. Antescofo: Anticipatory synchronization and control of interactive parameters in computer music. In *Proc. of the International Computer Music Conference*, 2008.

[8] S. Flossmann, M. Grachten, and G. Widmer. *Guide to Computing for Expressive Music Performance*, chapter Expressive Performance Rendering with Probabilistic Models, pages 75–98. Springer Verlag, 2013.

[9] K. P. Murphy. Fitting a conditional linear gaussian distribution. Technical report, University of British Columbia, 1998.

[10] D. Pelleg and A. W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proc. of the Seventeenth International Conference on Machine Learning*, pages 727–734, 2000.

[11] C. Raphael. Music plus one and machine learning. In *Proc. of the 27th International Conference on Machine Learning*, pages 21–28, 2010.

[12] M. Sarkar. Tablanet: a real-time online musical collaboration system for indian percussion. Master's thesis, MIT, 2007.