MULTI-PLAYER MICROTIMING HUMANISATION USING A MULTIVARIATE MARKOV MODEL

Ryan Stables
DMT Lab,
Birmingham City University
Birmingham, UK
ryan.stables@bcu.ac.uk

Satoshi Endo
Institute for Information-Oriented Control
Technische Universität München,
Munich, Germany
s.endo@tum.de

Alan Wing
SYMON Lab,
School of Psychology,
Birmingham University,
Birmingham, UK
a.wing@bham.ac.uk

ABSTRACT

In this paper, we present a model for the modulation of multi-performer microtiming variation in musical groups. This is done using a multivariate Markov model, in which the relationship between players is modelled using an interdependence matrix (α) and a multidimensional state transition matrix (S). This method allows us to generate more natural sounding musical sequences due to the reduction of out-of-phase errors that occur in Gaussian pseudorandom and player-independent probabilistic models. We verify this using subjective listening tests, where we demonstrate that our multivariate model is able to outperform commonly used univariate models at producing human-like microtiming variability. Whilst the participants in our study judged the real time sequences performed by humans to be more natural than the proposed model, we were still able to achieve a mean score of 63.39% naturalness, suggesting microtiming interdependence between players captured in our model significantly enhances the humanisation of group musical sequences.

1. INTRODUCTION

In electronically produced music, humanisation algorithms are often applied to percussive sequences in order to create a more natural sounding expressive performance. This is particularly useful when access to performers or equipment is limited, as events can be programmed onto a quantised grid and then modulated by a music producer, without the requirement for human performance. This process is often applied during the point of music creation from within the digital audio workstation and allows for the incorporation of sampled or synthesised instruments into a piece of music.

One of the main issues with current humanisation systems is that they do not necessarily represent the expressivity exhibited by a human agent, thus the process requires further editing in order to achieve a natural approximation of a human musician. Furthermore, the systems are unable to model the characteristics of group performance when used in a multi-channel environment. These problems are namely due to the fact that the majority of existing humanisation systems modulate the onset locations and respective velocities of an event instantaneously, using a pseudorandom variate, selected from a Gaussian window. Therefore in simulated multi-player performance, phase error is often introduced between the channels. This can actually reduce the naturalness of the performance, rather than enhance it due to perceptually unrealistic cues, generated by multiple instances of the algorithm running in parallel.

1.1. Modelling Microtiming

In this study, we focus specifically on extracting and modulating microtiming offsets in musical performance, this can be defined as the subtraction of an event at time n from a corresponding point on a reference track, as illustrated in Figure 1. Here, the reference grid represents a metronome running in parallel with the performed musical sequence. The challenge of the humanisation algorithm is to then estimate the distribution at $n + 1$, written as $P(\theta_{n+1})$. This is usually done independently of all other events in the sequence, based on a distribution centred around the n^{th} grid point, characterised by the parameters μ and σ.

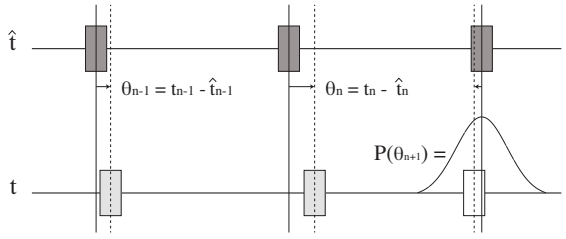

Figure 1: *Representation of a player, following a metronome. The offset measurements from the metronome are shown as θ_n, where $\hat{t}_n$ is the n^{th} metronomic event and t_n is the n^{th} event performed by player 1.*

Attempts have been made in previous studies to increase the naturalness of single-player humanisation systems by incorporating some form of intelligent processing into the variate generation procedure. In [1] for example, fuzzy logic has been used to model strike velocity deviation in a humanisation system, based on subjective rules, derived from domain knowledge. Similarly, microtiming deviation has been modelled using a number of different supervised machine learning techniques by [2]. These techniques are then used to apply the derived microtiming models to quantised sequences, in which they conclude the systems used to model percussive sequences significantly outperform the quantised version, when evaluated for natural expressivity. Microtiming for Brazillian Samba music is also estimated in [3] and [4], using a model based on the extraction of quarter-note patterns using K-Means clustering. Here, it is shown that the degree of expressive timing can be attributed to specific metrical positions, with examples in Samba music. This kind of information is omitted when pseudo-random models are applied, due to the variables being independently distributed for each event.

In previous work by Stables ([5], [6]), it has been shown that the process of stochastic humanisation can be improved using probabilistic temporal models to modulate a quantised sequence, based on microtiming measurements taken from professional musicians. Here, independently distributed variates $(P(\theta_{n+1}))$ were replaced by variates that were conditionally dependent in time $(P(\theta_{n+1}|\theta_n))$. In these studies it was shown that the measured sequences exhibited temporal patterns which could be synthesised using finite state machines. In both cases, the empirically developed models were shown to subjectively outperform quantised and Gaussian sequences for both perceived naturalness and musicality.

2. GENERATIVE MULTIVARIATE MODEL

Whilst the models described in section 1.1 work particularly well with single-player sequences, phase error is still introduced in mutichannel performance models due to the lack of inter-performer dependence. This means that when a probabilistic humanisation algorithm is applied to more than one track in a given session, extensive manual correction is often required in order to create a sense of cohesion between the separate channels. It is therefore necessary to consider ways in which a group of musicians can be modelled in parallel, thus preserving the inter-performer timing characteristics of a musical group.

If we make the assumption that the performed musical signals are stylised stochastic processes (as in studies such as [7] and [8]), we can use a Markov chain to estimate a transition through a discrete state-space $Z = \{z_1, z_2, \ldots, z_K\}$, where z_n represents the n^{th} state of the system, providing the sequence being modelled, satisfies the Markov property given in Eq. 1.

$$\begin{aligned} &P(\theta_{n+1} = i_{n+1}|\theta_0 = i_0, \theta_1 = i_1, \ldots, \theta_n = i_n) \\ &= P(\theta_{n+1} = i_{n+1}|\theta_n = i_n) \end{aligned} \tag{1}$$

Here, θ_n represents the n^{th} event and i_n represents the corresponding state. Each state in the model can be described using canonical form representation, consisting of a binary vector of length K, where $\sum_{k=1}^{K} \theta_k = 1$ and $\theta_k \in \{0, 1\}$. For example, in a 5-state model, if the n^{th} event is equal to z_3, we can use the representation $\theta_n = \{0, 0, 1, 0, 0\}^T$. This allows us to define a single-player model using Eq. 2.

$$P(\theta_{n+1}) = S\theta_n \tag{2}$$

Here, S is a state transition matrix (STM), representing the probability of a transition from $\theta_n = i_n$ to $\theta_{n+1} = i_{n+1}$ for $n = \{1, 2, \ldots, N\}$, where N is the number of events in the sequence. We then consider $P(\theta_n)$ to be the Probability Density Function (PDF) representation of θ_n. The canonical form of θ_{n+1} is then calculated using a rejection sampling technique, given here in Eq. 4.

$$(\theta_n)_i = \begin{cases} 1, & i = \beta \\ 0, & i \neq \beta \end{cases} \tag{3}$$

$$\beta = \begin{cases} \gamma_{1,k}, & [\gamma_{1,k}, \gamma_{2,k}] \in P(\theta_n) \\ repeat, & [\gamma_{1,k}, \gamma_{2,k}] \notin P(\theta_n) \end{cases} \tag{4}$$

Where $\gamma_{1,k}$ and $\gamma_{2,k}$ are pair-wise stochastic variables, evaluated against the n^{th} state distribution and β is the state vector index. For situations such as grouped musical performance, in which there are two or more conditionally dependent sequences, we can use a Multivariate Markov Chain (MVMC) model. This consists of the univariate model, estimated across M sequences being performed concurrently, weighted by some measure of inter-player dependence, given in Eq. 5.

$$P(\theta_{n+1}^{(i)}) = \sum_{k=1}^{M} \alpha_{i,k} S^{(i,k)} \theta_n^{(k)} \tag{5}$$

In the multivariate model, $\theta_n^{(i)}$ represents the state distribution of stream i in canonical form and the matrix $S^{(i,k)}$ gives the probability of a transition from the n^{th} state in stream i, to the $(n+1)^{th}$ state in stream k, as demonstrated in Eq. 6. When $k = i$, S represents a standard univariate STM. The weights $(\alpha_{i,k})$ in the model represent the interdependence factor between streams i and k, which can be derived empirically.

2.1. Pulse Approximation

As demonstrated in Figure 1, the estimation of microtiming parameters in the current model relies on an isochronous grid $(\hat{t})$ in order to calculate differentials $(\theta^{(i)})$ at any point in time (n). In single-player streams this model works particularly well if a player has performed the sequence to a click-track as we can use a metronomic grid to approximate $\hat{t}$. However due to the nature of group performance, it is relatively unlikely that the individual performers will follow the same click track, unless the musicians are independently contributing material to the musical piece. This trait is very common in multitrack recording, but less common in group performance. Using a metronomic model, we can represent the grid using Eq. 8.

$$\begin{aligned} &\theta_n^{(i)} = t_n^{(i)} - \hat{t}_n \\ &where, \quad \hat{t}_n = (n-1)\left(\tfrac{60}{\tau}\right) \end{aligned} \tag{8}$$

Where τ represents a measurement of fixed tempo and $t^{(i)}$ is the event generated by the i^{th} performer. In order to adapt this method for group performance, we need to estimate a global representation of tempo within the musical group. We can provide a simplistic model for this by taking the mean of the beat-spacings within each bar, across all players using Eq. 9.

$$S^{(i,k)} = \left\{ \begin{array}{cccc} p(\theta^{(i)} = z_1|\theta^{(k)} = z_1) & p(\theta^{(i)} = z_2|\theta^{(k)} = z_1) & \dots & p(\theta^{(i)} = z_K|\theta^{(k)} = z_1) \\ p(\theta^{(i)} = z_1|\theta^{(k)} = z_2) & p(\theta^{(i)} = z_2|\theta^{(k)} = z_2) & \dots & p(\theta^{(i)} = z_K|\theta^{(k)} = z_2) \\ \vdots & \vdots & \vdots & \vdots \\ p(\theta^{(i)} = z_1|\theta^{(k)} = z_K) & p(\theta^{(i)} = z_2|\theta^{(k)} = z_K) & \dots & p(\theta^{(i)} = z_K|\theta^{(k)} = z_K) \end{array} \right\} \tag{6}$$

$$i = \{1, 2, \dots, M\}, \quad k = \{1, 2, \dots, M\} \tag{7}$$

$$\hat{\tau}_m = \frac{1}{MB} \sum_{i=1}^{M} \sum_{n=1}^{B} (t_n^{'})^{(i)} - (t_{n-1}^{'})^{(i)} \tag{9}$$

Where $t_n^{'}$ represents an event that falls on a beat location and B is the number of beats in the bar. $\hat{\tau}_m$ then represents the estimated tempo for the m^{th} bar. This is now an estimated dynamic measurement of temporal drift and is updated each time a new bar is performed. The micro timing offsets are then subtracted from this grid, using the technique defined in Eq. 8, replacing τ with $\hat{\tau}_m$ and interpolated for n.

2.2. Inter-Player Dependence

We model the interdependence ($\alpha_{i,j}$) amongst performers in the group using lagged cross-correlation, in which player i's stream is lagged by a pre-defined number of events (n) and correlated with the stream of player j. This allows us to estimate the amount of dependence that one player has on another. This technique has been demonstrated by [9] to be optimal at a single event, suggesting that players are highly receptive to short-term variations in accompaniment. This measurement is demonstrated in Eq 10.

$$\alpha_{i,j} = \frac{1}{N} \sum_{k=0}^{N-n-1} \theta_k^{(i)} \theta_{k+n}^{(j)} \tag{10}$$

Where n is a non-negative integer representing the number of events to lag, set to 1 for this application.

3. EXPERIMENT: STRING QUARTET MODELLING

In order to evaluate the performance of the model, we analyse a professional quartet performing an excerpt from the 4^{th} movement of Hayden's String Quartet Op. 74 No. 1 in C-Major, the score for which is given in Figure 2. The quartet consisted of two violins, a viola and a cello, and the excerpt was chosen due to the number of notes being performed concurrently. The quartet have around 12 years experience performing together, and were shown by [9] to follow the lead violin player relatively closely. The excerpt, consisting of 12 bars was performed and recorded 15 times using the same equipment and the musicians were asked to perform using their natural expression. In total, each take contained 48 musical events, all of which were being performed by all members of the quartet at the same metrical positions in the bar.

Each player was recorded using an individual instrument microphone (DPA 4061), positioned on the body of each instrument with a rubber mount in order to reduce bleed in the recordings. The onsets from each player were then extracted using a spectral-flux based technique, and adjusted manually to improve accuracy. To find the microtiming offsets, the pulse was estimated at the beginning of each bar using the method defined in Eq. 9 and the events were subtracted using the technique defined in Eq. 8. The mean tempo for the recordings was found to be 105.0 BPM, with a standard deviation of 6.49. Figure 3 illustrates offset measurements from all 15 takes, with the mean of the results represented in black. Here, deviations are shown across all four performers playing concurrently.

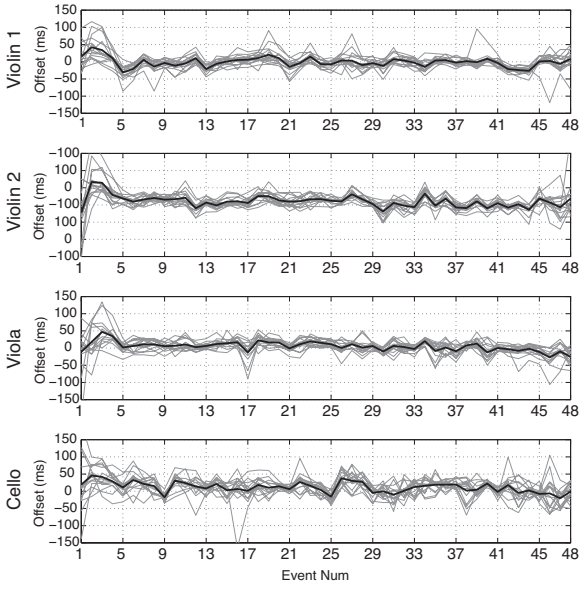

Figure 3: *A graphical representation of the microtiming deviation (θ) for all four performers. the measurements are taken across 15 takes of the same piece with the mean offset indicated in black, the vertical lines represent bar divisions.*

3.1. Subjective Evaluation

To evaluate the perceived naturalness of the model, subjective listening tests were conducted using a MUSHRA-based methodology [10]. The subjects were asked to rank each of the samples with a multi-stimulus interface and provide a rating between 0-100 for how naturally expressive each sample was perceived to be. Participants were informed that the experiments were based on professional musicians and were played an excerpt from a string quartet (not included in the stimuli) before the test began. In total 20 people participated in the experiment, all of whom were all aged between 18-35 and had normal hearing. All participants had some experience in performing or producing music.

The stimuli consisted of 25 versions of the same synthesised polyphonic sequence, the score for which was taken from Haydn's

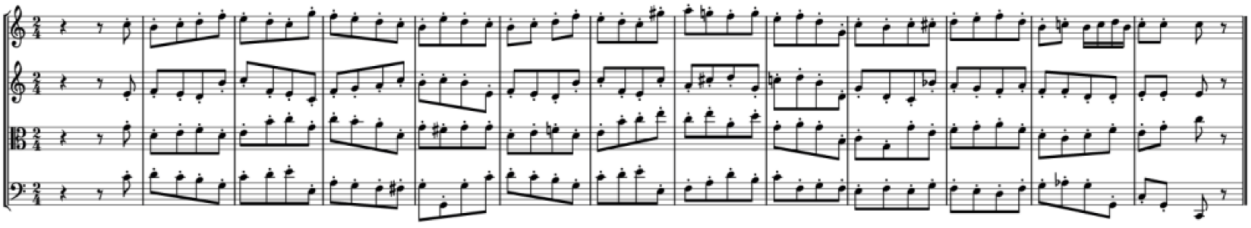

Figure 2: *The score of the excerpt taken from Hayden's Quartet Op. 74 No. 1 in C Major, in which 4 separate instrument parts are shown.*

Quartet Op. 74 and synthesised using a string ensemble pre-set from the Logic Studio 9 plug-in: EXS24 (Apple, CA, USA). The sequences were compiled by generating MIDI note-on messages and importing them into a sequencer. The MIDI was generated using 5 different techniques, these can be categorised as follows[1].

- *Quantised*: The note-on messages were quantised to a fixed grid, thus exhibiting no temporal variation.
- *Gaussian*: Each of the note-on messages were modulated using an independent Gaussian window.
- *MC*: The note-on messages for each channel were modulated using a conditionally independent Markov chain.
- *MVMC*: The note-on messages are modulated using the MVMC model presented in Eq. 5.
- *Human*: The onsets are taken from a dataset of human performers.

In order to isolate microtiming deviation, other parameters such as note-off and velocity were fixed to constant variables. The length of each event was fixed to 1/4-note length and the global tempo was varied across samples, bounded by measurements from the dataset. To control the mean and variance of the micro timing deviations across conditions, the μ and σ parameters used to characterise the distributions in the Gaussian method were derived from the dataset of human performers. This meant that all techniques were able to produce a similar range of θ values.

4. RESULTS

4.1. Performance Analysis

From our observations of a string quartet performing 15 iterations of a 12-bar of a piece in 4/4, we can identify characteristics of the musical group by performing analysis on the data. Firstly, the maximum microtiming deviation was measured to be 198.02ms and the minimum was -202.48ms. Overall the mean was 6.51ms, with a SD of 2.65ms. As the mean tempo was observed to be 105BPM, in 4/4 time signature, the maximum deviation was around 35.4% and the mean deviation was around 1.2% of the inter-onset interval (IOI).

The dependencies between each performer in the group are summarised in Eq. 11 and also shown using boxplots in Figure 4. Both of these diagrams represent the variable α in the model.

$$\alpha = \left\{ \begin{array}{cccc} 0.410 & 0.009 & 0.026 & 0.001 \\ 0.177 & 0.257 & 0.113 & 0.147 \\ 0.217 & 0.203 & 0.320 & 0.175 \\ 0.007 & 0.151 & 0.072 & 0.181 \end{array} \right\} \quad (11)$$

[1]Stimuli can be found at http://www.ryanstables.co.uk/data/dafx14

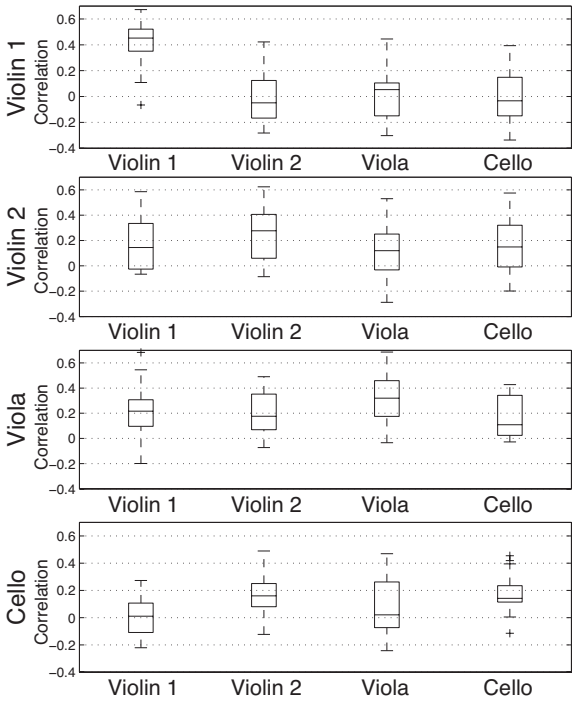

Figure 4: *Boxplot representation of inter-player dependence measured over 15 takes. This is measured using a lagged cross-correlation function.*

Here it is evident that the most highly correlated measurements taken from the data are based on lagged autocorrelation. This promotes the use of Markov chains in musical performance modelling as it suggests there is a strong relationship between an event (x_n) and it's predecessor (x_{n-1}) within the same stream. Generally, the 1^{st} violin has very low correlation scores with the other musicians in the group with a mean of 0.012 and a very high auto-correlation measurement. This suggests that they have adopted the role of lead performer. The other musicians in the group are generally more positively correlated with each other. Here, both the 2nd Violin and the viola player are following the lead violin, whilst the Cello is following the 2nd violin. We can calculate a leadership metric (l_α) for each player by taking the column-wise means, excluding the autocorrelation measurements at cell α_{ij} where $i = j$. This is illustrated in Eq. 12.

$$l_\alpha = \left\{ \begin{array}{cccc} 0.1337 & 0.1210 & 0.0703 & 0.1077 \end{array} \right\} \quad (12)$$

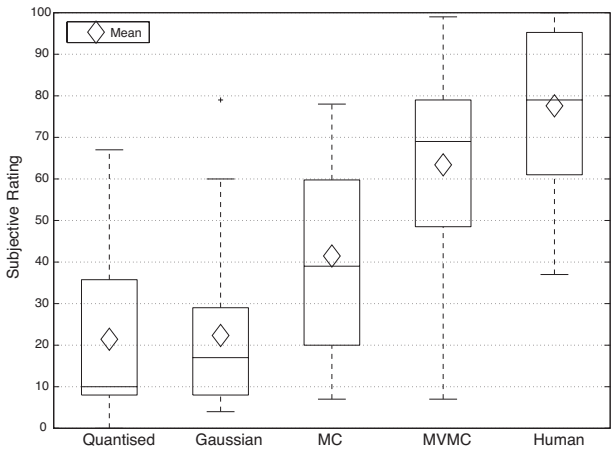

Figure 5: *A boxplot showing subjective listening test results taken from 20 subjects. The stimuli consisted of 5 samples taken from 5 categories (25 in total).*

Here, it is evident that the 1st Violin has the highest degree of leadership, reinforcing the suggestion that the performer has a leading role within the group.

4.2. Model Evaluation

In order to evaluate the naturalness of the model, we performed subjective tests to identify the similarity between the generated sequences and the performed sequences. The results from the subjective tests are illustrated in Figure 5. Here, it is evident that the microtiming sequences sampled from real musicians performed better than any of the synthetic samples with a mean score of 77.59%. The lowest scoring categories were Gaussian and Quantised models, which scored 22.33% and 21.42% respectively. The samples that were generated using the proposed multivariate model scored relatively highly with 63.39%, this was 21.94% higher than the closest category, which was the univariate model suggested in [5]. This result shows that the multivariate model performs slightly less favourably than using onsets taken directly from human performers, however it outperforms all existing methods for univariate modulation.

5. DISCUSSION

5.1. Model Performance

From the analysis of the string quartet, it is evident that the performers all seem to have stronger lagged autocorrelation scores (α_{ij}, where $i = j$), than cross-correlation scores ($i \neq j$). This would suggest that the internal representation of time held by each player takes priority over the external timings of group performance. Whilst these autocorrelation scores are significantly higher than the cross-correlation measurements, the performers still produce a sufficient amount of microtiming offset to cause potentially audible phase errors in the piece. This suggests the model's dependence matrix (α) is a significant factor as both the univariate model and the normally distributed model (with equivalent μ and σ parameters) underperform at producing timing sequences with natural expressivity. Subjectively, the multivariate model tends to produce much more confluent sequences than any of the univariate models running in parallel across multiple channels.

Whilst the subjective listening tests show an increased mean score for the multivariate model, suggesting the model is able to produce realistic musical sequences, there is a much higher variance than in other categories. This means there is uncertainty within the results, with some participants rating the system as low as 7/100. This is acceptable to an extent due to the relative uncertainty in the human samples, however it suggests there is room for improvement due to the inconsistency in results.

5.2. Implementation

Whilst we have demonstrated that the univariate models running in parallel do not perform particularly well for this application, the model allows for the conversion between univariate and multivariate methods by converting α to an identity matrix, imposing conditional independence on all streams. Similarly, we can alter the dependencies in α to change the characteristics of the musical group. If for example, the performance requires the group to closely follow Violin 1, the values in column 1 can be incremented, thus increasing the performers' leadership score (l_α). From an implementation standpoint, this is relatively simple to parameterise as users of the system can input values into the dependence matrix directly or via some mapping function.

Another key aspect to producing natural sounding rhythmic performance is tempo variation. In our listening tests, this was based on existing templates taken from our dataset. In most humanisation systems, this is ignored as control is generally maintained by the host application. For systems that wish to include this attribute, another variable can be added directly to the sum in Eq. 5, derived using the technique defined in Eq. 8. In the performances measured for this study, the tempo variation has a particularly high standard deviation due to the expressive nature of the music. In other genres such as pop-music, this may not be as important due to the prominence of quantisation and click-tracks.

6. CONCLUSION

In this paper, we have presented a model for the synchronous modulation of multiple streams of onsets using a multivariate Markov model. The model derives parameters from a user-defined corpus of multi-performer musical data and probabilistically applies modulation to a group of concurrent sequences. We can estimate the inter-player dependencies using lagged cross-correlation metric and approximate the pulse of the group using the bar-wise mean of all performers. The model is designed to alleviate the phase issues that arise when humanisation algorithms are applied to multiple sequences simultaneously.

We have demonstrated that the model outperforms univariate techniques including an instantaneous pseudorandom model and a Markov chain model applied independently to multiple channels, using data from a string quartet performing Haydn's Quartet Op. 74 No. 1 in C-Major. Through subjective listening tests, we observed an improvement of 21.94% accuracy on the closest synthesized category when measured for naturalness of expression. Whilst this was a significant improvement, sequences derived directly from human agents were still perceived to be more expressive than the model, indicating the importance and complexity of

the interdependence in multi-player musical performance that requires further attention.

7. REFERENCES

[1] L. O'Sullivan and F. Boland, "Towards a Fuzzy Logic Approach To Drum Pattern Humanisation," in *Proc. of the 13th Itl. Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, Sept. 19-21, 2010.

[2] M. Wright and E. Berdahl, "Towards machine learning of expressive microtiming in Brazilian drumming," in *International Computer Music Conference*, 2006.

[3] Fabien Gouyon, "Microtiming in samba de roda - preliminary experiments with polyphonic audio," in *Simpósio da Sociedade Brasileira de Computação Musical*, 2007.

[4] Luiz Alberto Naveda, Fabien Gouyon, Carlos Guedes, and Marc Leman, "Multidimensional microtiming in samba music," in *12th Brazilian Symposium on Computer Music*. SBCM, 2009, pp. 1–12.

[5] Ryan Stables, Jamie Bullock, and Ian Williams, "Perceptually relevant models for articulation in synthesised drum patterns," in *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[6] Ryan Stables, Cham Athwal, and Rob Cade, "Drum pattern humanization using a recursive bayesian framework," in *Audio Engineering Society Convention 133*, Oct 2012.

[7] Kevin Jones, "Compositional applications of stochastic processes," *Computer Music Journal*, vol. 5, no. 2, pp. 45–61, 1981.

[8] M. Kaliakatsos-Papakostas, M.G. Epitropakis, and M.N. Vrahatis, "Weighted markov chain model for musical composer identification," in *Applications of Evolutionary Computation*. 2011, vol. 6625 of *Lecture Notes in Computer Science*, p. 334–343, Springer.

[9] Alan M Wing, Satoshi Endo, Adrian Bradbury, and Dirk Vorberg, "Optimal feedback correction in string quartet synchronization," *Journal of The Royal Society Interface*, vol. 11, no. 93, pp. 20131125, 2014.

[10] ITURBS Recommendation, "1534-1: Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.