MULTIDIMENSIONAL SIMILARITY MODELLING OF COMPLEX DRUM LOOPS USING THE GROOVETOOLBOX

Fred Bruford[1] **Olivier Lartillot**[2] **SKoT McDonald**[3] **Mark Sandler**[1]
[1] Centre for Digital Music, Queen Mary University, United Kingdom
[2] RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo, Norway
[3] inMusic Brands inc, United States
fred.bruford@gmail.com, olivier.lartillot@imv.uio.no, smcdonald@inmusicbrands.com, mark.sandler@qmul.ac.uk

ABSTRACT

The *GrooveToolbox* is a new Python toolbox implementing various algorithms, new and pre-existing, for the analysis and comparison of symbolic drum loops, including rhythm features, similarity metrics and microtiming features. As part of the *GrooveToolbox* we introduce two new metrics of rhythm similarity and four features for describing the significant properties of microtiming deviations in drum loops. Based on a two-part perceptual evaluation, we show these four new microtiming features can each correlate to similarity perception, and be used with rhythm similarity metrics to improve personalized similarity models for drum loops. A new measure of structural rhythmic similarity is also shown to correlate more strongly to similarity perception of drum loops than the more commonly used Hamming distance. These results point to the potential application of the *GrooveToolbox* and its new features in drum loop analysis for intelligent music production tools. The *GrooveToolbox* may be found at: `https://github.com/fredbru/GrooveToolbox`

1. INTRODUCTION

Growing attention has been drawn to the applications of Music Information Retrieval (MIR) within the realm of music creation to improve upon conventional workflows and enhance creativity [13]. Due to their popularity in contemporary music, research into the analysis of drum loops is a field with strong potential to provide genuine value in real-world music production applications.

The problem of similarity modelling is a key element of this research. The ability to compare drum loops according to perceptually relevant qualities is an essential enabling factor in many plausible systems, such as drum loop recommendation systems, automatic drum loop generation systems and interfaces for navigating drum loop libraries.

For the purposes of this paper, one example use case for drum loop similarity modelling is to enable intelligent drum loop searching tools within *BFD3*, a virtual drum kit plugin [8]. *BFD3* generates realistic drum sounds based on audio renderings of expressive and unquantized symbolic sequences recorded by real drummers on an electronic drum kit. Including third-party expansions, over 7000 of these symbolic loops are available, providing rich potential for intelligent navigation or recommendation tools.

In **Section 2**, we give an overview of work related to intelligent drum production tools (IDPTs) and discuss possible improvements to their methods of drum loop analysis. In **Section 3** we introduce the *GrooveToolbox*, a Python toolbox primarily aimed towards use in drum loop analysis research. It contains implementations of many existing rhythm features and similarity measures for the analysis and comparison of symbolic drum loops with fixed tempo and metre. New algorithms are also provided: models of rhythmic structural similarity, and models of microtiming accounting for timing styles and mixtures of metrical subdivisions. This section gives an overview of the algorithms implemented in the *GrooveToolbox*.

In **Section 4**, we investigate the effectiveness of the algorithms contained within the *GrooveToolbox* via application to modelling similarity for drum loops. We test the effectiveness of our new polyphonic rhythm similarity measures against the commonly used Hamming distance, and test to what extent high-level rhythm feature-based similarity models can be improved with low-level rhythm similarity metrics and microtiming features to build individualized predictive models that could be used in user-aware IDPTs. In **Section 5** we summarize conclusions of this work, and detail further work required on the *GrooveToolbox* itself and in drum loop analysis generally.

2. BACKGROUND

2.1 Intelligent tools for drum loop production

The primary application of the *GrooveToolbox* is towards research into IDPTs. Much of this centres on the symbolic level rather than audio, and it largely centres on two applications: automatic generation of drum loops, and intelligent interfaces for exploring libraries of drum loops.

In general the goals of automatic drum loop generation in a music production context can be both to speed up the production process and help stimulate new ideas in the producer [17]. Much of the work in this area aims to generate variations on existing patterns to help producers create evolving drum parts. Genetic algorithms have been applied to achieve this, with the target vector derived from similarity metrics like the Hamming distance [21] or a feature set [16]. Also using genetic algorithms, [15] presents a system that interpolates between two existing drum patterns via a feature space. Variational autoencoders have also been used, generating loop variations that adapt to fit structural changes in an existing song [32] or to fit within a musical trio alongside melody and bass instruments [23].

A second area is in intelligent interfaces for exploring drum loop libraries. The design of intelligent user interfaces for improved music collection exploration is a well researched area with many successful applications [18]. Similarly, the mapping of drum loops on a 2D space via a similarity measure or dimensionality reduction of a feature space has potential to enable improved navigation of a large library. In [3], the authors map a large library of drum loops via the Self-Organizing Map, using a modified version of the Hamming distance as the similarity measure. In [11], the authors present a continuous, generative 2D space for drum patterns based on applying Multidimensional Scaling (MDS) to a set of rhythm features.

2.2 Improving drum loop analysis

In the IDPTs described above, symbolic drum loops are analyzed using rhythm features, such as density and syncopation, or rhythmic similarity measures such as the Hamming distance [30]. These may not capture all the important characteristics of complex loops. Two possible areas of improvement, informed by recent musicological research, are in the analysis of microtiming and rhythmic structure.

2.2.1 Microtiming

Microtiming can be defined as sub-rhythmic quasi-random or systematic timing deviations from a metrical grid in human performance. Representations of rhythms that are fit to a metrical grid are a requirement for many rhythm similarity measures, such as the Hamming distance, or features relying on metrical profiles like syncopation [20]. However, fitting rhythms to a grid removes microtiming information, which can be musically significant.

The timing 'feel' or 'groove' of a performance may be an important perceptual factor of drum loops; it has been shown that drummers can control the 'pushed' or 'laid-back' feel of their performance [5]. In [25], timing strategies in drumming for 'laid-back', 'ontop' and 'pushed' styles are measured for a group of drummers based on the typical back-beat rhythmic structure. It was found that their strategies can be formalized as specific timing interactions occurring on downbeat metrical positions. These are between the kick or snare and the hi-hat, or the metrical grid (or metronome) when there is no hi-hat present. The reference to hi-hat as well as grid was based on the understanding that when present the hi-hat usually acts as the time-keeper of the pattern. Detecting these timing interactions may be important for analysing groove in human-performed (or human-imitating) drum loops.

Secondly, gridded representations of rhythms do not account for swung rhythms, or loops where multiple subdivisions of a beat occur. Calculating a pattern-based rhythm similarity metric may still be desirable in these cases however. For example, a swung and unswung version of the same rhythm will be somewhat similar. Or, a loop in 4/4 time may have one instance of a triplet rhythm, in a fill for example. By incorporating a measure of deviation outside of a grid, we can manage these cases, whilst keeping the metrical reference required by other rhythm features.

2.2.2 Rhythmic Structure

The Hamming distance for rhythm similarity works by stepping through each metrical position in two rhythms, and counting the distance (difference) as the number of instances where rhythms contain different values (one a rest and the other an onset) in the same positions. Hence, in **Figure 1**, the Hamming distance would be 3. This measure is adapted to variable dynamics by using the difference in intensity or velocity as a weighting factor.

Though possibly the most popular rhythm similarity metric, the Hamming distance's stepwise nature fails to pick up regional rhythmic similarities. Onsets in similar but non-identical positions do not register as similar in the Hamming distance, even though perceptually they may be.

A measure of structural similarity may pick this up but requires the derivation of a structural representation. The recent rhythmic transformation model of [26] provides a means of doing so. In [26], an algorithm is described for characterizing rhythms as combinations of three types of ornamentation: syncopation, pickup (anacrusis) and density. Each of these is classified in terms of its position within a metrical profile and surrounding onsets. Syncopations are onsets placed in weak metrical positions, not followed immediately by an onset in a stronger position. Density ornamentations are placed in weak positions between two events in stronger positions. Pickups are placed in weak positions but followed immediately by an onset on a stronger position. Any rhythm can be defined as an ordered combination of these ornamentations against metrical profile, and any rhythm can be decomposed to a 'metronome' - an onset on each downbeat - in the same manner by reversing (removing) these ornamentations. This decomposition process can be used to simplify rhythms in a musicologically sound manner, where the simplified rhythmic representation is analogous to a structure.

3. GROOVETOOLBOX

GrooveToolbox is a Python toolbox for analysing and comparing rhythmic and microtiming qualities of drum loops in various formats. The toolbox contains functions for a variety of pre-existing features and new ones that account for rhythmic structure and microtiming. They fall within three groups: rhythm features, microtiming features and similarity measures. As the toolbox is designed to work with loops of fixed tempo, it does not contain features for tempo tracking or conventional metre detection. Nor does it provide timbral analysis as it works on a symbolic level.

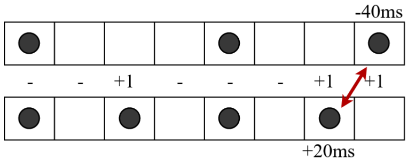


Figure 1. Measuring similarity of two short rhythms. Hamming distance = 3. With 16th note steps at 120BPM fuzzy Hamming distance = 2.26

Two Python toolboxes in the public domain relate to the *GrooveToolbox*. The *Rhythm Toolbox* [10] presented as part of [9] provides a starting point for drum loop analysis, with functions for two different types of syncopation features plus density features. However the feature set is limited, and it only supports MIDI files. The *GrooveToolbox* adds many algorithms to this set. It also supports BFD3 [8] format Groove files as an alternative to MIDI, MIREX format [1] and audio files through the integrated *ADTLib* drum transcription library [29].

The *SynPy* toolbox [28] also provides related functionality, consisting of implementations of seven syncopation models. While useful, the toolkit is designed for use on monophonic rhythms, and as such do not immediately apply to drum loops. In the *GrooveToolbox*, we implement one syncopation model [20] also found in *SynPy*, and add another designed specifically for drum patterns [34].

The algorithms implemented in the *GrooveToolbox* are collected from a range of research, with the aim of enabling comprehensive modelling of the perceptual qualities of drum loops. For re-implemented algorithms, we chose those which were experimentally verified as perceptually relevant and ensured they could account for variable onset velocities. The algorithms currently provided in the *GrooveToolbox* are listed in **Table 1**. In this section, we will describe the two new similarity models and four microtiming features. More details may be found at: `https://github.com/fredbru/GrooveToolbox`.

3.1 New Rhythm Similarity Models

3.1.1 Fuzzy Hamming distance

The fuzzy Hamming distance extends the Hamming distance by incorporating one metrical step of displacement along with the microtiming deviations of each onset. Where there is an onset in one rhythm but not the other, the algorithm looks ahead one step to look for a nearby onset. An instance of this is shown in **Figure 1**. If there is a nearby onset, the distance is reduced depending on how close that onset is. The microtiming deviations are also considered when two onsets occur at the same metrical position. In each case the timing difference between the two nearby onsets is incorporated in the similarity calculation. If present, the difference in onset velocity may also be used as a weighting factor as with the Hamming distance.

For the example of **Figure 1**, the value 1 in position 3 is the same as for the Hamming distance. The final two positions add 1 each to the Hamming distance, but the fuzzy Hamming counts the first of these as the timing difference between the two onsets, divided by the time of two metrical steps. At 120BPM, one 16th note lasts 125ms, so the similarity at this position $= (125 - 20 - 40)/250 = 0.26$.

With this proximity accounted for, the final position is calculated as usual, adding 1 to the distance. The overall distance is therefore 2.26 - close to the Hamming distance (3) but scored as more similar due to the proximity of the last two onsets. In a different case where the microtiming deviations in the two nearby onsets were removed, the fuzzy Hamming distance would be 2.50, higher to reflect the increased distance between the two onsets.

By accounting for possible similarity between nearby onsets, this method reduces the Hamming distance's limitation in detecting regional similarities. It also accounts for rhythms with microtiming deviations, such as swung rhythms, by not discarding them. Accounting for microtiming differences between onsets in the same position may also capture the overall difference in microtiming feel.

3.1.2 Structural Similarity

The structural similarity metric measures the similarity of a structural representation of two loops, derived following [26]'s transformation model (see **Section 2.2.2**). First, we remove any 'ghost notes', below a loudness threshold. Ornamentations are then found and reversed (removed) until any onsets only occur on downbeats. This results in representations of rhythmic structure at the downbeat level upon which the Hamming distance can be calculated.

3.2 Microtiming features

To develop features that describe the perceptual properties of microtiming deviations in drum loops, the first stage is to represent them in a form from which features can be extracted. In the context of drum pattern analysis, a sparse matrix format has been used by [12] to express microtiming. Here a matrix shows the timing deviation from the grid in milliseconds, positive (behind the beat) or negative (in front of the beat), for each onset. **Figure 2** shows this for a simple 2 beat pattern. The features extracted from this representation measure two types of microtiming effect: swing or metrical feel and performance styles.

	1	2	3	4	5	6	7	8
K	5	-	-	-	-	-	-	-
S	-	-	-	-	21	-	-	-
HH	-4	-	-8	-	-10	-	3	-

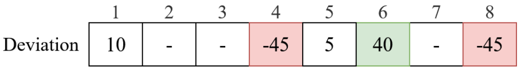
Figure 2. Matrix representation of timing deviations (ms) from 16th note positions in a 2 beat 120BPM kick, snare and hihat pattern (Laid-back event highlighted).

3.2.1 Swing and metrical feel

Using a sparse matrix representation of microtiming deviations alongside a rhythmic representation, swing can be detected along with the presence of triplets in a quadruple-time pattern. Swung onsets are detected as significantly delayed second eighth-notes, approximating the typically

Type	Feature Name	Description
Syncopation	Monophonic	Comparing onset pattern with hierarchical metrical profile [20]
	Polyphonic	Comparing interaction between instruments with metrical profile [34]
	Weak-Strong Ratio	Number of onsets not occurring on downbeats vs on downbeats [15]
Density	Absolute Density	Total onsets divided by number of possible onsets for any number of parts
	Relative Density	Density of one part divided by total density (cf. hiness feature [9])
	Syncopation Density	Syncopation divided by total number of onsets (cf. hisyness feature [9])
Complexity	Rhythmic Complexity	Quadratic mean of density and syncopation [27]
Periodicity	Autocorrelation Skewness	Skewness of autocorrelation curve [22]
	Autocorrelation Max Amplitude	Maximum amplitude of autocorrelation curve [22]
	Harmonicity	Harmonicity of autocorrelation curve, (primarily pulse clarity) [19]
	Symmetry	Proportion of onsets at the same position in 1st and 2nd half of pattern. [22]
Intensity	Average	Mean of all velocity values in pattern [15]
	Standard Deviation	Standard deviation of all velocity values in pattern [15]
Swing and metrical feel	**Swing-ness**	Whether loop is swung, weighted by number of swung notes
	Triplet-ness	Whether loop contains any triplets, weighted by number of triplet notes
Microtiming style	**Laidback-ness**	Microtiming style as number of push/laid-back events
	Timing accuracy	Mean of absolute timing deviation from grid of all onsets
Similarity measures	Hamming distance	Counting number of metrical positions where values (onset/rest)‘ are different
	Fuzzy Hamming distance	Hamming with 1 step lookahead, distance weighted by microtiming
	Structural similarity	Similarity of patterns simplified using [26]'s transformation algorithm

Table 1. List of features and similarity measures currently implemented in the *GrooveToolbox*. New features are in **bold**.



	1	2	3	4	5	6	7	8
Deviation	10	-	-	-45	5	40	-	-45

Figure 3. Matrix representation of timing deviations (ms) for 16th note metrical positions in a 2 beat 120BPM rhythm. Red = swung events, green = triplet events.

understood 2:1 eighth-note swing ratio. Although musically these are considered as eighth notes, they fall into sixteenth note positions when quantized, with significant negative (ahead of the position) deviations. The ‘swing-ness’ feature first records whether these timing deviations occur or not, returning 0 for no swing or 1 for swing. This is then weighted by the number of swung onsets to model perceptual salience of the swing. The deviation threshold for swing is calculated dependent on the tempo of the rhythm. The ‘triplet-ness’ feature is calculated in the same way, but also records the second note of a triplet as significantly delayed second eighth note. A triplet note detected from a microtiming matrix is shown in green in **Figure 3**.

3.2.2 *Microtiming style*

Overall timing accuracy is calculated as the mean of all absolute timing deviations per onset in a loop. For onsets classed as swung or triplets, the deviation is calculated from the ‘ideal’ triplet or swung note position.

Following the timing interaction classification of [25] as described in **Section 2.2.1** the laidback-ness feature counts the number of laid-back timing events subtracted from pushed events, with a negative score meaning an overall ‘pushed’ loop and positive a ‘laid-back’ loop. Thus a pattern's feel is calculated based on the detection of specific timing discrepancies above a perceptual threshold, known to impart a given feel from drumming performance analysis. We modified [25] by also counting for ride cymbal in place of hihat due to their similar musical roles.

Based on analysis of timing accuracy in BFD3's library, we chose a threshold of 12ms as a one that would disregard performance noise. However, ideally this would be calculated per drummer as in [25]. An example ‘laid-back’ pattern is shown in **Figure 2**. The highlighted event is ‘laid-back’, as there is a discrepancy on the downbeat between (in this case) snare and hihat that is above the threshold. For this pattern, the timing accuracy value would be 8.5, and laidback-ness 1.

4. EVALUATION

To evaluate our new algorithms and address open questions in drum loop analysis, a two-part experiment was carried out into modelling similarity for BFD3's drum loops, using perceptual data from humans collected via listening test. Three research questions were addressed:

1. *Should models of similarity for drum loops rely on rhythm similarity metrics, feature sets or both?*
   One approach to modelling drum loop similarity is to adapt rhythm similarity metrics [30] that measure distances between onset patterns, used for example in [3, 21, 31]. An alternative is to model similarity as a combination of higher-level rhythm features [11, 16]. While the two are not usually combined, both may be important as they emphasize different information.

2. *Can the models of microtiming proposed be used in modelling similarity of drum loops?*
   Existing IDPTs, described in **Section 2.1**, tend to assume simpler quantized and unswung rhythms. To apply this work to complex loops that are unquantized or swung, or include multiple subdivisions of the metre, models that can account for microtiming deviations could be important.

3. *How do the new rhythm similarity models compare to the Hamming distance?*
   We investigate alternate ways of measuring rhythmic similarity by testing the two new rhythm similarity measures proposed against the Hamming distance.

4.1 Data collection

The dataset consisted of similarity ratings for 80 pairs of BFD3's drum loops (160 total) as provided by 21 participants in a listening test, first collected in [4]. Loops were generated through the same virtual kit, BFD3's 70s Rock kit, chosen due to its generic timbre. Loops were collated equally from 8 genre groups: Blues/Country, Rock, Metal, Jazz, Funk, Reggae/Latin, Pop and Dance/Hiphop. Tempo, metre and loop length were fixed at 120bpm, 2 bars and 4/4 time. Some were swung, and a small number contained triplet rhythms. The test was distributed online using the Web Audio Evaluation Tool [14], with participants representing a range of musical and technical experience.

The listening test used a pairwise comparison methodology. Participants rated how similar two loops were on a continuous scale with five equally spaced markers (*Completely Different, Different, Slightly Different/Slightly Similar, Similar, Identical*). This was to maximize the number of loops in the study, whilst keeping the test length reasonable (30-40 minutes), ensuring multi-genre validity but giving more fine similarity ratings than in the common triadic comparison test design [2]. 5 training pairs were given at the start of the test, and 10 pairs were repeated at the start and end to test participants' internal consistency. The inter-rater reliability (IRR) calculated for all raters across all comparisons using the intraclass correlation coefficient (ICC) in (2,1) form, was 0.73, a moderate-to-good agreement. This is expected, as it is known that musical similarity perception is very individualized [33], with low IRR a challenge in musical similarity studies [7]. The average internal consistency of all participants, calculated as the median ICC (2,1) between ratings for the 10 repeated pairs, was 0.85, equal to good consistency.

4.2 Evaluation Design

Based on this perceptual data, our evaluation was formed in two parts to address the three research questions. The evaluation was designed with consideration of the individualized nature of similarity perception.

4.2.1 Overall perceptual relevance of new models

First we evaluated the new similarity and microtiming models based on their correlation to listeners' similarity ratings. The extent to which the models relate to perceived similarity was measured as the Pearson correlation between the feature score and median similarity rating of the 21 participants. The spread of ratings per pair was approximately normally distributed (D'Agostino-Pearson test $p >$ 0.01) so the median across raters is used. While this indicates if the features proposed relate to overall perceived similarity, limited IRR means that more precise analysis of the performance of these features against an average of ratings may not be valid. The second part of the experiment therefore uses individuals' ratings instead.

4.2.2 Building Individualized Similarity Models

In this part, the third research question is addressed. This experiment will find to what extent rhythm feature-based models can be adapted for use in complex drum loops when combined with microtiming features, and whether a combination of rhythm similarity metric and feature set can offer a better similarity model than either alone can. The aim here is to build similarity models that are predictive, investigating the utility of the models in practical use-cases that may require precise, fine-grained models. Due to low IRR, this requires we develop models for individuals' ratings separately, echoing the concept of 'user-aware' MIR [24]. The models are tested against individual ratings of 7 participants, chosen as those with highest internal consistency (for repeated pairs median ICC = 0.92), and whose ratings were normally distributed (D'Agostino and Pearson test $p > 0.01$), meaning regression models were applicable. Not all participants' individual ratings fit a normal distribution, and not all had high enough internal consistency for precise prediction, so not all could be used.

For each participant we tested seven feature combinations: rhythm features, microtiming features, the best similarity metric from **Section 4.3.1** and each combination of the three. We used all features in **Table 1**. The density features were calculated across three instrument groups separately: low (kick), mid (snare and toms) and high (cymbals) as in [11]. All other features were calculated with all instruments combined. Fitting a regression model to the 7 participants for 7 conditions, we measured the predictive power of the models as the explained variance (R-squared) for each feature combination and participant.

For the single similarity measure we used linear regression. For the feature sets, Partial Least-Squares (PLS) regression was chosen [35]. This was chosen because the features exhibit a high degree of colinearity, due to the large number of features and the inherent co-dependence of musical qualities. This combined with the high predictor-cases ratio means that linear modelling does not work. PLS regression combines elements from multiple linear and principal component regression, and has been found to work well in multidimensional musical emotion prediction [6]. Its use of principal component analysis to create latent prediction variables alleviates the problems of colinearity and high predictor-cases ratio, but limits interpretability of the significance of specific features. This is partly why we evaluate features differently in **Section 4.3.1**. To choose the best number of principal components for each model consistently, we used the Bayesian Information Criterion (BIC), which accounts for overfitting by penalising model complexity against model performance.

4.3 Results & Discussion

4.3.1 Overall perceptual relevance of new models

The correlation between the features and median similarity ratings is shown in **Table 2**. All models exhibited statistically significant correlation ($p < 0.05$).

The structural similarity measure exhibits higher correlation to the median similarity ratings than the ordinary velocity-weighted Hamming distance (t=1.69, p=0.046). It could therefore be the case for complex drum loops that listeners compare rhythms at a more global structural level

Feature/Model	Correlation *r*	*p*
Hamming Distance	0.59	9.7e-9
Structural Similarity	0.65	6.1e-11
Fuzzy Hamming Distance	0.56	6.1e-8
Swing-ness	0.46	1.9e-5
Triplet-ness	0.49	3.4e-6
Timing accuracy	0.33	2.9e-3
Laidback-ness	0.22	0.046

Table 2. Pearson *r* and *p-value* between median similarity ratings and model difference values.

rather than a precise low level. Alternatively, the structural representation could indicate shared genre between loops by showing the approximate locations of events. Given this seemingly strong performance, using the structural similarity measure over Hamming distance may be advantageous.

For the laidback-ness feature, an issue was that values were typically low, as there is often not a significant difference in microtiming styles between loops. Based on the 12ms deviation threshold, in only 21 of the 80 comparisons was there a difference between timing styles, with the remaining three quarters of the comparisons returning 0. Looking at the correlation between the laidback-ness feature and similarity ratings just for these 21 comparisons where timing style is different, correlation is stronger (r = 0.47, p = 0.0033). One interpretation is that a feature like this, which models a precise low-level quality, is only relevant for a similarity comparison when there is a significant difference in this quality. While further investigation is required, this may point to the use of adaptive feature weightings for similarity comparisons that select features based on the relevance to a particular comparison.

The other microtiming features exhibited moderate to good correlation, with the swing-ness and triplet-ness features being approximately the same. This may be because loops in our dataset with onsets matching to second triplet positions likely have notes in swung positions too, so there is little practical difference in their values. For a larger dataset this difference may be more significant. The fuzzy Hamming distance did not differ significantly from the regular Hamming distance. The correlations of both the swing and microtiming style features indicate that a better way to incorporate microtiming in similarity models could be as a separate set of features modelling global characteristics of microtiming, rather than being inserted into rhythmic similarity measures. Overall, it appears that these features are able to some extent to capture perceptually salient features of microtiming deviations in drum loops.

4.3.2 Building Individualized Similarity Models

The results of this part are shown in **Figure 4**. Comparing the median r^2 score for the 7 participants, it can be seen that the combination of rhythm and microtiming features with structural similarity measure results in the best predictive model (r^2 = 0.56), closely followed by rhythm feature and microtiming model (r^2 = 0.51). This confirms that both microtiming and structural similarity models can improve rhythm feature-based similarity models in this case. However, there is still improvement required before they can accurately predict similarity perception.

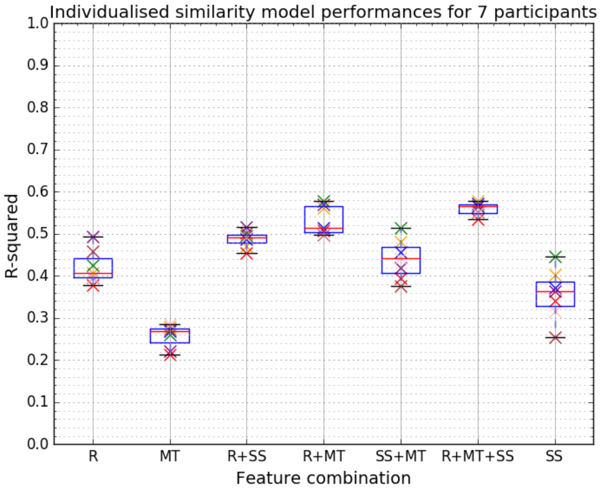


Figure 4. Model performance as R-squared value for combinations of rhythm **R** and microtiming **MT** feature sets with structural similarity feature **SS** for each participant.

There are a few possible reasons for this. As mentioned in **Section 4.3.1**, the system of deriving a fixed weighting of features in a multidimensional similarity model may not be the best way to model similarity; instead, adaptive weighting schemes may be required that weigh features according to their relevance in a given comparison. While the feature set seems comprehensive, there may be some qualities of drum loops that are not effectively being modelled, in particular features that explicitly detect style or genre. Similarity ratings from more listeners should be collected in the future to construct further personalized models and verify these findings for a wider range of listeners.

5. CONCLUSIONS & FURTHER WORK

We presented a new toolbox for drum loop analysis, with implementations of pre-existing algorithms and new ones for analysing and comparing rhythmic structure and microtiming. These were found to correlate to perceived similarity of drum loops. The rhythmic structural similarity metric was shown to correlate at least as well as the conventional Hamming distance to similarity perception in drum loops. It has been shown that the ideal model of similarity for complex drum loops combines rhythm and microtiming features with a rhythm similarity metric. These results all have implications in future work on IDPTs.

As further work, new algorithms should be developed to improve similarity models in the *GrooveToolbox*, in particular ones that explicitly model stylistic similarity. Here more investigation into rhythmic grouping or structure could be useful. Due to the complex nature of similarity perception it is difficult to infer the practical utility of similarity models from this evaluation. For a more ecologically valid understanding, we will next evaluate them in the context of an IDPT. An approach to combining features that accounts for the possibly attention-based nature of similarity perception could also be a valuable direction. Given the added requirement for personalized models, the next step is to investigate methods such as active learning to learn an adaptive similarity model from a user.

6. ACKNOWLEDGEMENTS

This work has been funded in part by the UK Engineering and Physical Sciences Research Council (EPSRC) and by a Short-Term Scientific Mission grant provided by Nordic SMC. This work was also partially supported by the Research Council of Norway through its Centers of Excellence scheme, project number 262762, the TIME project, grant number 249817 and the MIRAGE project, grant number 287152. We would like to thank Anne Danielsen and the members of the TIME project at the University of Oslo for their valuable feedback and advice at the early stages of this work.

7. REFERENCES

[1] 2019:Drum Transcription - MIREX Wiki. `https://www.music-ir.org/mirex/wiki/2019:Drum_Transcription`. Accessed May 2020.

[2] Hamish Allan, Daniel Müllensiefen, and Geraint A. Wiggins. Methodological considerations in studies of musical similarity. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2007.

[3] Fred Bruford, Mathieu Barthet, SKoT McDonald, and Mark Sandler. Groove Explorer: An Intelligent Visual Interface for Drum Loop Library Navigation. In *IUI Workshops*, 2019.

[4] Fred Bruford, Mathieu Barthet, SKoT McDonald, and Mark Sandler. Modelling Musical Similarity for Drum Patterns: A Perceptual Evaluation. In *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, 2019.

[5] Anne Danielsen, Carl Haakon Waadeland, Henrik G Sundt, and Maria AG Witek. Effects of instructed timing and tempo on snare drum sound in drum kit performance. *The Journal of the Acoustical Society of America*, 138(4):2301–2316, 2015.

[6] Tuomas Eerola, Olivier Lartillot, and Petri Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2009.

[7] Arthur Flexer and Taric Lallai. Can we increase inter- and intra-rater agreement in modelling general music similarity? *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[8] FXpansion. BFD3 `https://www.fxpansion.com/products/bfd3/`, Accessed May 2020.

[9] Daniel Gómez Marín. *Similarity and style in electronic dance music drum rhythms*. PhD thesis, Universitat Pompeu Fabra.

[10] Daniel Gómez Marín. Rhythmtoolbox `https://github.com/danielgomezmarin/rhythmtoolbox`, Accessed May 2020.

[11] Daniel Gómez-Marín, Sergi Jorda, and Perfecto Herrera. Drum rhythm spaces: from global models to style-specific maps. In *International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2017.

[12] Kahl Hellmer and Guy Madison. Quantifying microtiming patterning and variability in drum kit recordings: A method and some data. *Music Perception*, 33(2):147–162, 2015.

[13] Eric J. Humphrey, Douglas Turnbull, and Tom Collins. A brief review of creative MIR. In *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 2013.

[14] Nicholas Jillings, David Moffat, Brecht De Man, and Joshua D. Reiss. Web Audio Evaluation Tool: A browser-based listening test environment. In *12th Sound and Music Computing Conference*, 2015.

[15] Maximos Kaliakatsos-Papakostas. Generating drum rhythms through data-driven conceptual blending of features and genetic algorithms. In *International Conference on Computational Intelligence in Music, Sound, Art and Design*, pages 145–160. Springer, 2018.

[16] Maximos A. Kaliakatsos–Papakostas, Andreas Floros, and Michael N. Vrahatis. evoDrummer: Deriving Rhythmic Patterns through Interactive Genetic Algorithms. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design*, pages 25–36. Springer Berlin Heidelberg, 2013.

[17] Peter Knees, Kristina Andersen, Sergi Jordà, Michael Hlatky, Günter Geiger, Wulf Gaebele, and Roman Kaurson. Giantsteps-progress towards developing intelligent and collaborative interfaces for music production and performance. In *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*, pages 1–4. IEEE, 2015.

[18] Peter Knees, Markus Schedl, and Masataka Goto. Intelligent user interfaces for music discovery: The past 20 years and what's to come. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[19] Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, and Jose Fornari. Multi-feature modeling of pulse clarity: Design, validation and optimization. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2008.

[20] H. C. Longuet-Higgins and C. S. Lee. The Rhythmic Interpretation of Monophonic Music. *Music Perception: An Interdisciplinary Journal*, 1(4):424–441, 1984.

[21] Cárthach O. Nuanáin, Perfecto Herrera, and Sergi Jorda. Target-based rhythmic pattern generation and

variation with genetic algorithms. In *Sound and Music Computing Conference*, 2015.

[22] Maria Panteli, Bruno Rocha, Niels Bogaards, and Aline Honingh. A model for rhythm and timbre similarity in electronic dance music. *Musicae Scientiae*, 1:24, 2016.

[23] Adam Roberts, Jesse Engel, and Douglas Eck. Hierarchical variational autoencoders for music. In *NIPS Workshop on Machine Learning for Creativity and Design*, 2017.

[24] Markus Schedl, Arthur Flexer, and Julián Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539, 2013.

[25] George Sioros, Guilherme Schmidt Câmara, and Anne Danielsen. Mapping timing strategies in drum performance. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[26] George Sioros, Matthew EP Davies, and Carlos Guedes. A generative model for the characterization of musical rhythms. *Journal of New Music Research*, 47(2):114–128, 2018.

[27] Georgios Sioros and Carlos Guedes. Complexity driven recombination of midi loops. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[28] Chunyang Song, Marcus Pearce, and Christopher Harte. Synpy: a python toolkit for syncopation modelling. In *Sound and Music Computing Conference*, 2015.

[29] Carl Southall, Nicholas Jillings, Ryan Stables, and Jason Hockman. Adtweb - an open source browser based automatic drum transcription system. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[30] Godfried T. Toussaint. A comparison of rhythmic similarity measures. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2004.

[31] Richard Vogl, Matthias Leimeister, Cárthach Ó Nuanáin, Michael Hlatky, Sergi Jordà Puig, and Peter Knees. An intelligent interface for drum pattern variation and comparative evaluation of algorithms. *Journal of the audio engineering Society. 2016; 65 (7/8): 503-13.*, 2016.

[32] I-Chieh Wei, Chih-Wei Wu, and Li Su. Generating structured drum pattern using variational autoencoder and self-similarity matrix. *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[33] Geraint A Wiggins. Models of musical similarity. *Musicae Scientiae*, 11:315–338, 2007.

[34] Maria AG Witek, Eric F Clarke, Mikkel Wallentin, Morten L Kringelbach, and Peter Vuust. Syncopation, body-movement and pleasure in groove music. *PloS one*, 9(4), 2014.

[35] Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.