AN EVALUATION FRAMEWORK AND CASE STUDY FOR RHYTHMIC CONCATENATIVE SYNTHESIS

Cárthach Ó Nuanáin, Perfecto Herrera, Sergi Jordà
Music Technology Group
Universitat Pompeu Fabra
Barcelona
{carthach.onuanain, perfecto.herrera, sergi.jorda}@upf.edu

ABSTRACT

In this paper we present and report on a methodology for evaluating a creative MIR-based application of concatenative synthesis. After reviewing many existing applications of concatenative synthesis we have developed an application that specifically addresses loop-based rhythmic pattern generation. We describe how such a system could be evaluated with respect to its its objective retrieval performance and subjective responses of humans in a listener survey. Applying this evaluation strategy produced positive findings to help verify and validate the objectives of our system. We discuss the results of the evaluation and draw conclusions by contrasting the objective analysis with the subjective impressions of the users.

1. INTRODUCTION

MIR-based applications are becoming increasingly widespread in creative scenarios such as composition and performance [14] [7] [8]. This is commensurate with the prevalence of sampling-based approaches to sound generation, thus the desire is to develop more rich and descriptive understanding of the underlying content being used.

One of the primary difficulties faced with designing instruments for creative and compositional tasks remains the elaboration of an appropriate evaluation methodology. Indeed, this is a trending challenge facing many researchers [2], and numerous papers address this directly with various proposals for methodological frameworks, some drawing from the closely related field of HCI (Human Computer Interaction) [13], [16], [11]. More generally the evaluation of computer composition systems has also been the subject of much discussion in the literature. One frequent benchmark for evaluating algorithmic music systems is a type of Turing test where the success criterion is determined by the inability of human listener to discern between human and computer-generated music. As Hiraga [11] notes, however, these kind of tests can be problematic for two reasons. Firstly, it makes the assumption that the music generated by the algorithm is intended to sound like music produced by humans, rather than something to be treated differently. Secondly it ignores other facets of the system that imperatively needs evaluation, such as the interface and the *experience*. Pachet also finds issue with simplistic Turing test approaches to music evaluation [18]. He repeats, for instance, the view that unlike the traditional Turing test which evaluated the ability to synthesis believable natural language, no such "common-sense" knowledge exists for aspects of music.

We have designed and developed an MIR-driven instrument that uses concatenative synthesis to generate looped rhythmic material from existing content. In terms of evaluation we face the challenge of evaluating an MIR driven software system, thus subject to the same scrutiny facing any information retrieval system that needs to be appraised. We also face the challenge of evaluating the system as a musical composition system that needs to serve the composer and listener alike.

In the next section we will give the reader brief familiarity with the instrument in terms of its implementation and functionality. Subsequently, existing concatenative systems will be reported on in terms of their evaluation methodologies (if any). Section 3 will propose the evaluation framework in questions and the results will be reported. We will conclude the paper with our impressions on what we have learnt and scope for improvement in terms of the system itself and the evaluation methodology.

2. INSTRUMENT DESCRIPTION

Concatenative synthesis builds new sounds by combining together existing ones from a corpus. It is similar to granular synthesis differing only in the order of size of the grains: granular synthesis operates on microsound scales of 20-200ms whereas concatenative synthesis uses musically relevant unit sizes such as notes or even phrases. The process by which these sounds are selected for resynthesis is a fundamentally MIR-driven task. The corpus is defined by selecting sound samples, optionally segmenting them into onsets and extracting a chosen feature set to build descriptions of those sounds. New sounds can finally be synthesised by selecting sounds from the corpus according to

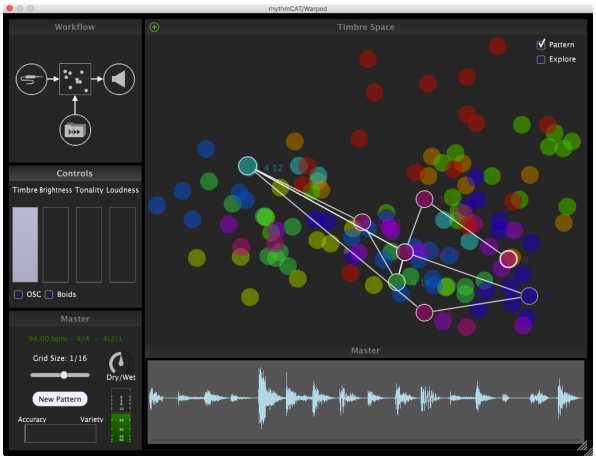

Figure 1: Instrument Interface

Author (Year)	Evaluation
Schwarz (2000)	No
Zils & Pachet (2001)	No
Hazel (2001)	No
Hoskinson & Pai (2001)	No
Xiang (2002)	No
Kobayashi (2003)	No
Cardle et al. (2003)	Videos of use cases
Lazier & Cook (2003)	No
Sturm (2004)	No
Casey (2005)	Retrieval Accuracy
Aucouturier & Pachet, (2005)	User Experiment
Simon et al. (2005)	Algorithm Performance
Jehan (2005)	Algorithmic evaluation
Schwarz (2005)	No
Weiss et al. (2009)	No
Frisson et al. (2010)	No
Hackbarth (2010)	No
Bernardes (2014)	Author's impressions

Table 1: Evaluation in Concatenative Systems

a unit selection algorithm and connecting them in series, maybe applying some cross-fading to smooth disparities in the process.

Concatenative synthesis has a long history of application in speech synthesis [15]. One of the most well-known works in the area of musical concatenative synthesis is CataRT [22] but there are many other systems referenced in the literature including some commercial implementations. Bernardes [3] provides a thorough summary of these based on similar reports in [25] and [21].

Our system (Figure 1) resembles many concatenative synthesis applications that offer a 2D timbre space for exploration. Where it distinguishes itself is in its sound generation strategy and mode of interaction for the user. Implemented as a VST plugin, it is an inherently loop-based instrument. It records and analyses incoming audio from the host as target segments according to a metrical level and concatenates units of sound from the corpus to generate new loops with varying degrees of similarity to the target loop. This similarity is determined by the unit selection algorithm, the central component in concatenative systems.

2.1 Unit Selection

The unit selection algorithm is quite straightforward. For each unit i in the segmented target sequence (e.g. 16-step) and each corpus unit j (typically many more), the concatenation *unit cost* $C_{i,j}$ is calculated by the weighted Euclidean distance of each feature k as given by Equation 1, where a and b are the values of the features in question.

$$C_{i,j} = \sqrt{\sum_{k=1}^{n} w_k (a_k - b_k)^2} \quad (1)$$

In terms of feature selection, after consulting a number of different articles [10], [20] and [27], dealing with feature extraction and rhythm we decided on a combination of MFCCs, loudness, spectral centroid and spectral flatness.

These unit costs are stored in similarity matrix M. Next we create a matrix M' of the indices of the ascendingly sorted elements of M. Finally a concatenated sequence can be generated by returning a vector of indices I from this sorted matrix and playing back the associated sound file. To retrieve the closest sequence V_0 one would only need to return the first row (Equation 3).

$$V_0 = (I_{0,i}, I_{0,i+1}..., I_{0,N}) \quad (2)$$

Returning sequence vectors solely based on the row restricts the possibilities to the number of rows in the matrix and is quite limited. We can extend the number of possibilities to i^{j-T} units if we define a similarity threshold T and return a random index between 0 and $j - T$ for each step i in the new sequence.

3. EVALUATION OF CONCATENATIVE SYNTHESIS

As we were researching existing works in the table presented by Bernardes, [3] we were struck by the absence of discussion regarding evaluation in most of the accompanying articles. This table we reproduce here (Table 1) amended and abridged with our details on the evaluation procedures (if any) that were carried out.

Some of the articles provided descriptions of use cases [4] or at least provided links to audio examples [24]. Notably, many of the articles [23], [9] consistently made references to the role of "user", but only one of those actually conducted a user experiment [1]. By no means is this intended to criticise the excellent work presented by these authors. Rather it is intended to highlight that although evaluation is not always an essential part of such experiments - especially in "one-off" designs for single users such as the author as composer - it is an underexplored aspect that could benefit from some contribution.

We can identify two key characteristics of our research that can inform what kind of evaluation can be carried out. Firstly it's a retrieval system, and can be analysed to determined its ability to retrieve relevant items correctly. Sec-

ondly it is a system that involves users or more precisely, musical users. How do we evaluate this crucial facet?

Coleman has identified and addressed the lack of subjective evaluation factors in concatenative synthesis [5]. In his doctoral thesis he devotes a chapter to a listening survey conducted to determine the quality of a number of different algorithmic components of the system under consideration. He asks the listeners to consider how well the harmony and timbre of the concatenated sequences are retained. In a previous paper [17] we conducted a similar-style listening survey to determine the ability of a genetic algorithm to create symbolic rhythmic patterns that also mimic a target input pattern. Evaluation strategies need to be tailored specifically for systems, but if the system is intended to retrieve items according to some similarity metric, and the material is musical, then a listening survey should be critical. Furthermore, and echoing Coleman's work, we would emphasise that whatever the application of a concatenative system, the evaluation of the *timbral* quality is essential.

In light of these elements we also devised a quantitative listening survey to examine musical output of the system not only in terms of its facility in matching the target content perceptually but also in producing musically pleasing and meaningful content.

4. METHOD

4.1 Evaluation and Experimental Design

Given the rhythmic characteristics of the system we formulated an experiment that evaluated its ability to generate new loops based on acoustic drum sounds. We gathered a dataset of 10 breakbeats ranging from 75 BPM to 142BPM and truncated them to single bar loops. Breakbeats are short drum solo sections from funk music records in the 1970s and exploited frequently as sample sources for hip-hop and electronic music. This practice has been of interest to the scientific community, as evident in work by Ravelli et al. [19], Hockman [12] and Collins [6].

In turn, each of these loops was then used as a seed loop for the system with the sound palette derived from the remaining 9 breakbeats. Four variations were then generated with 4 different distances to the target. These distances correspond to indices into the similarity matrix we alluded to in Section 2, which we normalise by dividing the index by the size of the table. The normalised distances then chosen were at 0.0 (the closest to the target), 1.0 (the furthest from the target) and two random distances in ranges 0.0 - 0.5 and 0.5 - 1.0.

Repeating this procedure 10 times for each target loop in the collection for each of the distance categories, we produced a total of 40 generated files to be compared with the target loop. Each step in the loop was labelled in terms of its drum content, for example the first step might have a kick and a hi-hat. Each segmented onset (a total of 126 audio samples) in the palette was similarly labelled with its corresponding drum sounds producing a total of 169 labelled sounds. The labellings we used were K = Kick, S = Snare, HH = Hi-hat, C = Cymbal and finally X when the content wasn't clear. Figure 2 shows the distribution of onsets by type in the full corpus of segmented units. Another useful statistic is highlighted in Figure 3, which plots the distribution of onsets for each step in the 16 step sequence for the predominant kick, snare and hi-hat for the 10 target loops. Natural trends are evident in these graphs, namely the concentration of the kick on the 1st beat, snares on the 2nd and 4th beat and hi-hats on off beats.

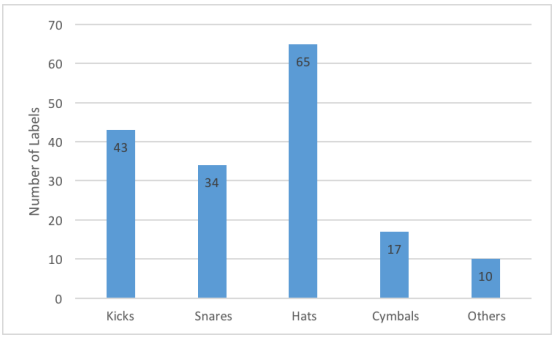

Figure 2: Distribution of Sounds in Corpus

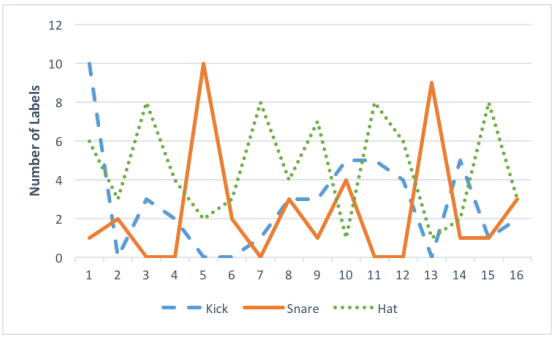

Figure 3: Distribution of Sounds in Target Loops

4.2 Retrieval Evaluation

The aim of the experiment was first to determine how well the system was able to retrieve similarly labelled "units" for each 1/16th step in the seed loop. To evaluate the ability of the algorithm to retrieve correctly labelled sounds in the generated loops we defined the accuracy A by equation 3, based on a similar approach presented in [26]. We make a simplification that corresponding HH and X and C labels also yield a match based on our observation that their noisy qualities are very similar, and some of the target loops used did not have onsets sounding at each 1/16th step.

$$A = \frac{number\ of\ correctly\ retrieved\ labels}{total\ number\ of\ labels\ in\ target\ loop} \quad (3)$$

4.2.1 Retrieval Results

By studying the Pearson's correlation between the retrieval ratings and the distance, we can confirm the tendency of smaller distances to produce more similar patterns by observing the moderate negative correlation (ρ = -0.516, p

<0.001) between increased distance and the accuracy ratings (Figure 4).

An interesting observation is that when we isolate the retrieval accuracy ratings to kick and snare we see this correlation increase sharply to (ρ = -0.826, p <0.001), as can be seen in Figure 5.

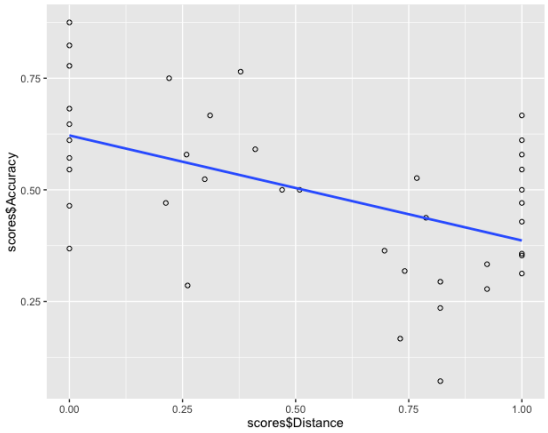

Figure 4: Scatter Plot and Regression Line of Retrieval Accuracy with Distance for all Drum Sounds

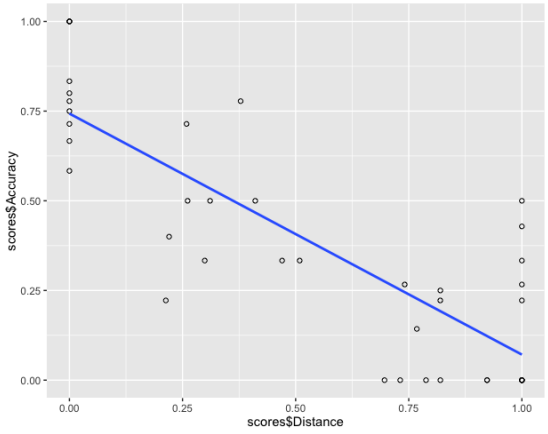

Figure 5: Scatter Plot and Regression Line of Retrieval Accuracy with Distance for Kick and Snare

Delving into the data further, we can identify 3 different categorical groupings that demonstrate predictable trends in terms of the central tendencies and descending retrieval accuracy (Figure 6). We label these categories A, B and C with the breakdown of the number of patterns and their corresponding distance ranges as follows:

- A - 10 patterns - 0.0
- B - 9 patterns - [0.2 - 0.5]
- C - 21 patterns - [0.5 - 1.0]

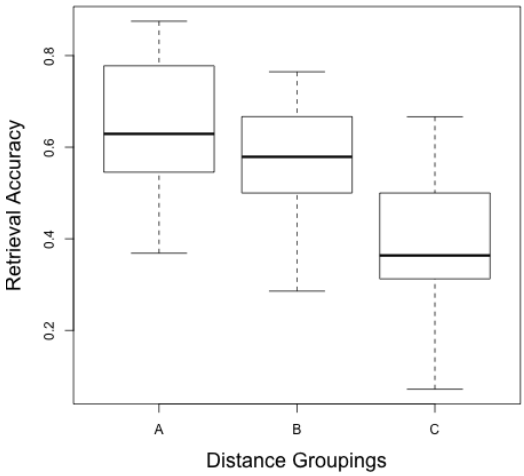

Figure 6: Central Tendencies of Retrieval Ratings for the Similarity/Distance Categories

4.3 Listener Evaluation

The retrieval accuracy gives the objective ratings of the system's capability for retrieving correctly labelled items. This information may not be consistent with the human listener's perceptual impression of similarity, nor does it give any indication whether the retrieved items are musically acceptable or pleasing. To assess the human experience of the sonic output and to compare with the objective ratings of the system, we conducted a listening survey which will be described here.

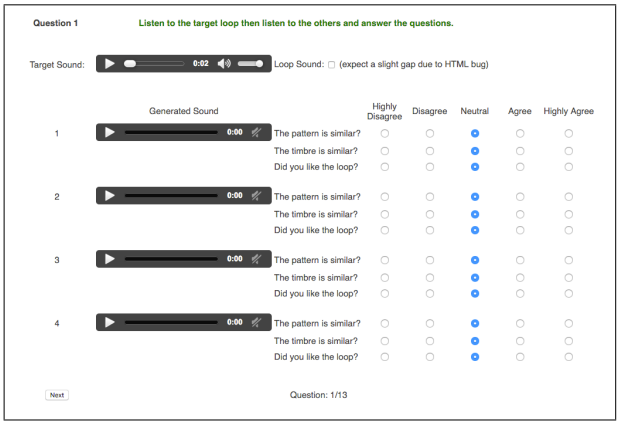

Figure 7: Screenshot of Web Survey

To directly compare and contrast with the retrieval evaluation the same 40 loops generated by the system and used in the retrieval analysis were transferred to the listening survey. The survey itself was web-based (Figure 7) and took roughly 15-20 minutes to complete. Participants were presented with the seed pattern and the generated patterns and could listen as many times as they liked. Using a 5 point Likert scale the participants were then asked to rate their agreement with the following statements:

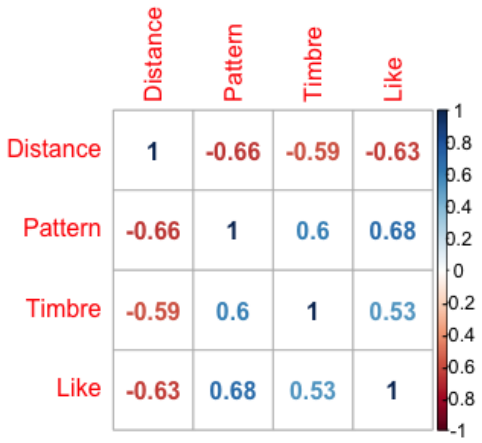

Figure 8: Correlations Between Distance and Subjective Ratings of Pattern Similarity, Timbre Similarity and Liking

- Is the rhythmic pattern similar?
- Is the timbre similar?
- Did you like the loop?

Twenty one participants in all took part in total, drawn from researchers at the authors' institute as well as friends and colleagues with an interest in music. Twenty out of the 21 participants declared they were able to play a musical instrument Ten of the 21 participants specified they played a percussion instrument and 9 reported they could read notation. In the instructions we provided brief explanations of the key terms and audio examples demonstrating contrasting rhythmic patterns and timbres.

4.3.1 Listener Evaluation Results

The survey data was analysed using Spearman's rank correlation on the mode of the participants' responses to each loop stimulus with the associated distance of that loop. We identified a moderate to strong negative correlation for each of the pattern, timbre and "likeness" aspects ($p < 0.01$ in all instances). This can be observed in the red values in the correlation matrix presented in Figure 8.

It should be evident that the subjective listening data conforms quite well to the findings of the objective retrieval rating. Increased distance resulted in decreased retrieval accuracy which in turn corresponded to a decrease in listener ratings for qualities pertaining to pattern similarity and impression of timbral similarity in the sounds themselves. Furthermore, it was revealed that the aesthetical judgement of the generated loops, encapsulated by the "likeness" factor, also followed the trend set out by the objective algorithm. We were curious to establish whether any particular subject did not conform to this preference for similar loops, but examining the individual correlation coefficients revealed all to be negative (all participants preferred more similar sounding patterns).

5. CONCLUSIONS

In this paper we presented a proposal for a framework that evaluates concatenative synthesis systems. Using a system that we developed which specifically generates rhythmic loops as a use case we demonstrated how such a framework could be applied in practice. An application-specific experiment was devised and the objective results and subjective results showed favourably the performance of the similarity algorithm involved. It is hoped that by providing a well-documented account of this process other researchers can be encouraged to adapt comparable evaluation strategies in creative applications of MIR such as concatenative synthesis.

6. ACKNOWLEDGMENTS

This research has been partially supported by the EU-funded GiantSteps project (FP7-ICT-2013-10 Grant agreement nr 610591). [1]

7. REFERENCES

[1] Jean-Julien Aucouturier and François Pachet. Ringomatic: A Real-Time Interactive Drummer Using Constraint-Satisfaction and Drum Sound Descriptors. *Proceedings of the International Conference on Music Information Retrieval*, pages 412–419, 2005.

[2] Jeronimo Barbosa, Joseph Malloch, Marcelo M. Wanderley, and Stéphane Huot. What does " Evaluation " mean for the NIME community? *NIME 2015 - 15th International Conference on New Interfaces for Musical Expression*, page 6, 2015.

[3] Gilberto Bernardes. *Composing Music by Selection: Content-based Algorithmic-Assisted Audio Composition*. PhD thesis, University of Porto, 2014.

[4] Mark Cardle, Stephen Brooks, and Peter Robinson. Audio and User Directed Sound Synthesis. *Proceedings of the International Computer Music Conference (ICMC)*, 2003.

[5] Graham Coleman. *Descriptor Control of Sound Transformations and Mosaicing Synthesis*. PhD thesis, Universitat Pompeu Fabra, 2015.

[6] Nick Collins. *Towards autonomous agents for live computer music: Realtime machine listening and interactive music systems*. PhD thesis, University of Cambridge, 2006.

[7] Matthew E. P. Davies, Philippe Hamel, Kazuyoshi Yoshii, and Masataka Goto. AutoMashUpper: An Automatic Multi-Song Mashup System. *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013*, pages 575—-580, 2013.

[1] http://www.giantsteps-project.eu

[8] Dimitri Diakopoulos, Owen Vallis, Jordan Hochenbaum, Jim Murphy, and Ajay Kapur. 21st century electronica: Mir techniques for classification and performance. In *International Society for Music Information Retrieval Conference*, pages 465–469, 2009.

[9] Benjamin Hackbarth. Audioguide : A Framework for Creative Exploration of Concatenative Sound Synthesis. *IRCAM Research Report*, 2011.

[10] Perfecto Herrera, Amaury Dehamel, and Fabien Gouyon. Automatic labeling of unpitched percussion sounds. In *Audio Engineering Society 114th Convention*, 2003.

[11] Rumi Hiraga, Roberto Bresin, Keiji Hirata, and Haruhiro Katayose. Rencon 2004: Turing Test for Musical Expression. *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 120–123, 2004.

[12] Jason A. Hockman and Matthew E. P. Davies. Computational Strategies for Breakbeat Classification and Resequencing in Hardcore, Jungle and Drum & Bass. In *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, pages 1–6, 2015.

[13] William Hsu and Marc Sosnick. Evaluating interactive music systems: An HCI approach. In *Proceedings of New Interfaces for Musical Expression*, pages 25–28, 2009.

[14] Eric J Humphrey, Douglas Turnbull, and Tom Collins. A brief review of creative MIR. *International Society for Music Information Retrieval*, 2013.

[15] Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1:373–376, 1996.

[16] Chris Kiefer, Nick Collins, and Geraldine Fitzpatrick. HCI Methodology For Evaluating Musical Controllers: A Case Study. *Proceedings of the 2008 International Conference on New Interfaces for Musical Expression (NIME-08)*, pages 87–90, 2008.

[17] Cárthach Ó Nuanáin, Perfecto Herrera, and Sergi Jorda. Target-Based Rhythmic Pattern Generation and Variation with Genetic Algorithms. In *Sound and Music Computing Conference 2015*, Maynooth, Ireland.

[18] François Pachet and Pierre Roy. (Manufac) Turing Tests for Music. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI 2015), Workshop on "Beyond the Turing Test*, 2015.

[19] Emmanuel Ravelli, Juan P. Bello, and Mark Sandler. Automatic rhythm modification of drum loops. *IEEE Signal Processing Letters*, 14(4):228–231, 2007.

[20] Pierre Roy, François Pachet, and Sergio Krakowski. Analytical Features for the classification of Percussive Sounds: the case of the pandeiro. In *10th Int. Conference on Digital Audio Effects (DAFx-07)*, pages 1–8, 2007.

[21] Diemo Schwarz. Current Research In Concatenative Sound Synthesis. In *Proceedings of the International Computer Music Conference*, pages 9–12, 2005.

[22] Diemo Schwarz, G Beller, B Verbrugghe, and S Britton. Real-Time Corpus-Based Concatenative Synthesis with CataRT. *Proceedings of the 9th International Conference on Digital Audio Effects*, pages 18–21, 2006.

[23] Ian Simon, Sumit Basu, David Salesin, and Maneesh Agrawala. Audio analogies: creating new music from an existing performance by concatenative synthesi. *Proceedings of the International Computer Music Conference*, 2005:65–72, 2005.

[24] Bob L. Sturm. Matconcat: An Application for Exploring Concatenative Sound Synthesis Using Matlab. In *7th International Conference On Digital Audio Effects (DAFx)*, pages 323–326, 2004.

[25] Bob L. Sturm. Adaptive Concatenative Sound Synthesis and Its Application to Micromontage Composition. *Computer Music Journal*, 30(4):46–66, 2006.

[26] Lucas Thompson, Simon Dixon, and Matthias Mauch. Drum Transcription via Classification of Bar-Level Rhythmic Patterns. In *International Society for Music Information Retrieval Conference*, pages 187–192, 2014.

[27] Adam Tindale, Ajay Kapur, George Tzanetakis, and Ichiro Fujinaga. Retrieval of percussion gestures using timbre classification techniques. *Proceedings of the International Conference on Music Information Retrieval*, pages 541–545, 2004.