Green Open Access Version - Originally published in J. Audio Eng. Soc., vol. 69, no. 1/2, pp. 40–53, (2021 Jan/Feb)

Manifold learning methods for visualization and browsing of drum machine samples

Jordie Shier JSHIER@UVIC.CA
AES Student Member
University of Victoria, Victoria, Canada

Kirk McNally KMCNALLY@UVIC.CA
AES Member
University of Victoria, Victoria, Canada

George Tzanetakis GTZAN@CS.UVIC.CA
AES Member
University of Victoria, Victoria, Canada

Ky Grace Brooks KY.BROOKS@PROTONMAIL.COM
AES Student Member
Center for Interdisciplinary Research in Music Media and Technology (CIRMMT)
School of Information Studies
McGill University, Montreal

Green Open Access Version. Please note that this manuscript is released under the Journal of Audio Engineering Society's "Green" OA option, required by NSERC, the funding body for this research. This manuscript is the final version accepted for publishing, but is not presented in the JAES graphical template. The original version, published in the JAES, is available here: `https://www.aes.org/e-lib/browse.cfm?elib=21015`.

Abstract

The use of electronic drum samples is widespread in contemporary music productions, with music producers having an unprecedented number of samples available to them. The task of organizing and selecting from these large collections can be challenging and time-consuming, which points to the need for improved methods for user interaction. This paper presents a system that computationally characterizes and organizes drum machine samples in two-dimensions based on sound similarity. The goal of the work is to support the development of intuitive drum sample browsing systems. The methodology presented explores time segmentation, which isolates temporal subsets from the input signal prior to audio feature extraction, as a technique for improving similarity calculations. Manifold learning techniques are compared and evaluated for dimensionality reduction tasks, and used to organize and visualize audio collections in two-dimensions. This methodology is evaluated using a combination of objective and subjective methods including audio classification tasks and a user listening study. Finally, we present an open-source audio plug-in developed using the JUCE software framework that incorporates the findings from this study into an application that can be used in the context of a music production environment.

1. INTRODUCTION

The first commercial electronic drum machine was released in 1959 by the Rudolph Wurlitzer Corporation. Marketed as an automatic rhythm accompaniment, the Side Man featured ten preset electronic drum sounds and twelve predefined rhythmic patterns. In recent decades the use of drum machine and drum samples has grown to such a degree that they are now ubiquitous in music productions, with contemporary music producers having an unprecedented number of drum samples available to them. The issue of navigating and selecting from large collections of audio samples has been expressed by expert users in the field of electronic dance music (EDM) production when asked about their desires for future technological advancements [1]. Intelligent music production (IMP), including automated systems intended to aid creativity and improve user workflow, is a growing area of research in the area of creative Music Information Retrieval (MIR) [2, 3] and the problem of automatic sorting, selection, and auditioning for large sample library collections within a music production context has not been fully studied.

The use of digital audio workstations (DAWs) in contemporary music production is widespread and the use of audio samples, including drum machine samples, is common practice within these tools. Although most mainstream DAWs support the navigation of audio samples directly within the software, samples are typically displayed in a list-based user-interface and any searching is based on filenames or semantic tagging [4].

The focus of our work is to explore different methods used to characterize kick and snare drum audio samples, sort them based on sound similarity, and visualize them in two-dimensions. This research is an extension of prior work by the authors [5, 6], where it was found that using short segments taken from full length audio samples increases the ability to computationally characterize and classify kick and snare drum samples.

The methodology presented here utilizes time segmentation as a pre-processing step to audio feature extraction, where small segments are first extracted from the full-length audio sample. 13 different time segment combinations of different sample lengths and sample start offsets are included for comparison. A novel method for selecting different time segments per audio feature in order to maximize variance is introduced, and achieves promising results. In order to visualize drum samples we explore and compare a variety of dimensionality reduction methods including the traditional linear Principle Component Analysis (PCA) method as well as other non-linear manifold learning approaches. PCA and audio classification tasks, performed on a database of 4228 kick and snare drum samples from 250 individual drum machines, is used to explore and quantitatively evaluate our approach. A listening user study was designed to develop ground truth measurements of sound similarity for a selection of kick and snare drum samples, and the results of this user study were used to evaluate the perceptual relevance of our proposed methodology.

The remainder of this paper is structured as follows: A review of related work is provided in Section 2. Sections 3, 4, and 5 outline the research performed which explores and evaluates the methods used for characterizing and visualizing kick and snare drum samples. Section 6 describes the listening test used, and the statistical analysis of the results.

2. RELATED WORK

There is increasing interest by practitioners of music production to adopt technology that improves creativity and aid user workflow [7], and there is a growing body of work in the fields of intelligent music production and MIR to address these needs. The development of tools to visualize audio collections and enhance the way users interact with these collections is an established area of research in MIR. The concept of visualization in MIR is comprehensively reviewed by Cooper et al. [8], and is updated by Schedl et al. to represent contemporary work [9]. AudioQuilt is a notable example of an experimental application for visualizing audio samples [10]. AudioQuilt uses metric learning and kernalized sorting algorithms to visualize audio samples in two dimensions. An application that utilizes similar techniques to the work presented here is DrumSpace, developed by Turquois et al. [11]. DrumSpace performs feature extraction on a set of drum sounds and displays them on a two dimension interface using a Student-t Stochastic Neighbour Embedding (t-SNE) algorithm. User participants responded positively to the two dimensional interface for exploring drum samples, however, reported being confused by the arrangement of samples when organized by sound similarity. Turquois et al. suggest the use of colour in visualizations as a potential approach to improving the perception of sound similarity. Another related application with a novel user interface is Mixploration [12], which utilizes a self-organizing map algorithm for visualization and manipulation of multiple audio mixing parameters on a two dimensional interface. Moving beyond two dimensions, experimental applications with novel user interfaces for exploring sound collections in three dimensions are reviewed by Tzanetakis and Cook [13]. A unique approach to sample retrieval is proposed by Knees and Andersen [4], in which users query audio using visual sketches of their mental images of that sound. A recent noteworthy commercial audio plugin called XO[1] developed by XLN Audio organizes user drum samples on a two dimension interface for sample exploration.

An essential component of the audio visualization tools introduced above is the extraction of data from the source material that is to be visualized. Computational audio analysis is at the core of many MIR applications and improving these techniques has been a focus of research within this field. Related audio analysis studies focussing on percussion sounds includes work by Herrera et al. [14], in which a set of features is used to analyze a large set of drum samples, including 33 different drum classes comprised of both acoustic and synthetic sources. In subsequent work, Herrera et al. focussed on analyzing un-pitched percussive sounds, including a mix of acoustic and synthetic kick and snare drums [15]. A unique component of their study is a classification test that classifies percussion instruments of the same type by their make or model. Several new music content descriptors related to percussion were later proposed by Herrera et al., including "percussivity index" which estimates the amount of percussion in a musical audio file as well as "percussion profile" which is an estimate of the balance between drums and cymbals in a percussion track [16]. Sound similarity and classification of electronic drum samples is a component of the experiments performed by Tutzer [17]. Focusing on classification of cymbal sounds, Souza et al. [18] achieved high accuracy scores using Support Vector Machines.

It is typical in computational audio analysis to examine the entire duration of a sound being studied, however there is evidence that emphasizing certain portions of an audio signal

[1] `https://www.xlnaudio.com/products/xo`

or considering only a temporal subset of the entire signal may have a beneficial effect on analysis in terms of accuracy to human perception. Toiviainen explored emphasizing the onset of an audio signal as a method to optimize auditory images, which are representations of an audio signal that are computed using an auditory model [19]. An auditory model takes into account filtering performed by the inner and outer ear, the dynamics of the basilar membrane, the mechanical response of hair cells, and the electrical response of auditory nerves. They found significant improvements in correlations between auditory images and subjective similarity ratings when the onset of a signal was emphasized. Building upon this work, Pampalk et al. focused on sound similarity between percussion sounds of the same class and found that using shorter sample lengths resulted in improvements in the characterization of kick drum sounds [20]. In their work, four distinct percussion categories were examined (kicks, snares, high toms, and low toms) and two different models of sound similarity were compared: the MPEG-7 model and aligned auditory images.

3. METHODS

3.1 Pre-Processing

Pre-processing is applied to all samples prior to audio analysis. All audio samples are down-mixed to mono and resampled to a rate of 44.1kHz if required. A normalization step includes applying ReplayGain[2].

3.2 Time Segmentation

In this step, a segment of the kick or snare sample is selected for further processing. For clarity we will refer to the segment extracted from the original as the sample, and the original kick or snare sample as the input signal. Choices for sample lengths are derived from work by Danielsen et al. [21], where a 23ms window is used to segment snare drum samples, a 50ms segment is used by Bello [22], a 100ms segment is used by Herrera et al. [16], and Pampalk et al. [20] compared three different sample lengths: the first 250ms, the first 1000ms, and the entire input signal. In this work, we use sample lengths of 25ms, 100ms, and 250ms, 500ms. We also consider three different starting positions for each sample length. The starting position is derived from the signal envelope and is selected as the point in time during the attack portion of the signal when the envelope reaches a certain percentage of the maximum amplitude. Choices for these positions are derived from the Essentia documentation[3], the toolbox used for audio feature analysis, and include 20%, 50%, and 90%. Therefore, there are 12 different combinations of sample length and sample start offset which we evaluate in this study. The entire sample duration with no start offset is also considered.

The sample time segmentation s is selected as follows:

1. Calculate the the signal envelope g from the input signal x by rectifying x and applying a non-symmetric low-pass filter:

[2] http://wiki.hydrogenaud.io/index.php?title=ReplayGain_1.0_specification

[3] www.essentia.upf.edu/documentation

$$g[n] = (1-k) * g[n-1] + k * |x[n]|$$

Where

$$k = \begin{cases} e^{\frac{-T_s}{a_t/1000}} & \text{for } |x[n]| > g[n-1] \\ e^{\frac{-T_s}{r_t/1000}} & \text{for } |x[n]| <= g[n-1] \end{cases}$$

$$T_s = \text{Sampling period}, a_t = 10\text{ms}, r_t = 1500\text{ms}$$

2. Determine the sample start position n_{start} where n_{start} is the first occurrence of n such that $g[n] = p * \max(g)$ and $p \in \{0.2, 0.5, 0.9\}$. Figure 1 shows an example of these start positions calculated on the envelope of a drum signal.

3. Calculate the sample end position s_{end} and extract the sample s from the full input signal x with length of $l \in \{25ms, 100ms, 250ms, 500ms\}$:

$$n_{end} = n_{start} + l/1000 * T_s$$
$$s = x[n_{start} : n_{end}]$$

Figure 2 shows an example of a time segmentation with a length of 100ms starting at 50% of the attack computed on a drum signal.

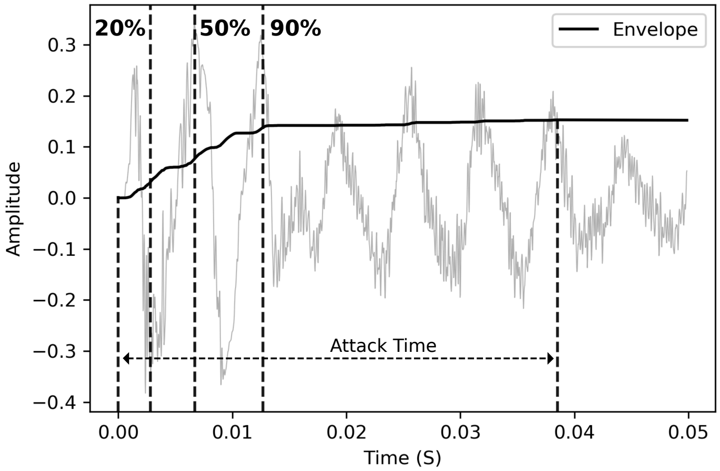


Figure 1: *Choices for sample position.* Choices are 20%, 50%, and 90% of the maximum amplitude of the signal envelope. The attack time is also shown and spans from 0% to 100% of the maximum amplitude of the signal envelope.

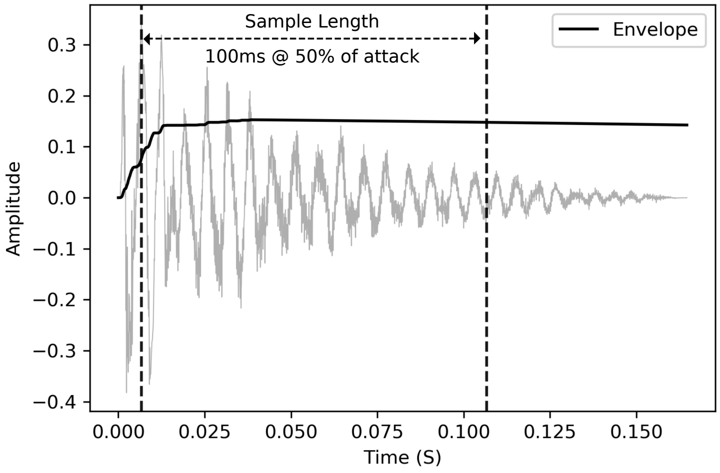


Figure 2: *Sample length and position.* A snare drum audio file with a sample length of 100ms positioned at 50% of the attack envelope.

3.3 Feature Extraction

Audio feature extraction was performed using the Essentia library [23]. This library is selected based on findings reported by Moffat et al. in their evaluation of audio feature extraction toolkits [24], where it was found to be the most comprehensive library with regards to feature coverage. The features selected for use are from those defined both within the ISO/IEC-defined MPEG-7 format, used in previous audio characterization work by Peeters et al. [25], and prior work into the classification of percussion sounds. These features are described in detail in subsequent work by Peeters et al. [26] and include MFCCs, HFC, spectral, and temporal features, as well as a set of 27 bark-scale frequency bands which together constitute a 133-dimensional feature-space. These features have been shown to work well across a wide variety of audio analysis and classification tasks and our analysis of time segmentations and features shows how they can be selected and adapted to the task at hand.

For computation of the spectral features, a 2048 sample Hann window using a hop-size of 1/8, derived from Pampalk et al. [20] is used. Calculations for each feature using the 2048 sample window are summarized over time using mean and standard deviation. An equal loudness filter is applied as a pre-processing step before calculating spectral centroid, kurtosis, skewness, and spread as suggested by the Essentia documentation[4], for details on the equal loudness filter see the ReplayGain specification[5]

3.4 Mixed Time Segmentation

A mixed segmentation method is introduced here as an experimental approach to using time segmentation in conjunction with feature extraction and dimensionality reduction. This method seeks to maximize variance in the dataset by independently selecting a time

[4] `https://essentia.upf.edu/reference/std_LowLevelSpectralEqloudExtractor.html`

[5] `http://wiki.hydrogenaud.io/index.php?title=ReplayGain_1.0_specification`

segmentation for each feature. This method is computed using the following procedure:

1. Let $\mathbf{D} \in \mathbb{R}^{n \times s \times d}$ be the dataset resulting from feature extraction where n is the number of kick or snare samples, s is the number of time segmentations computed (13), and d is the number of audio features extracted (133).

2. For each feature d_i, select the time segmentation s_{max_i} that results in the greatest variance for that feature.

3. Compile a new matrix $\mathbf{D}_{mixed} \in \mathbb{R}^{n \times d}$ such that each feature d_i was calculated using s_{max_i}.

3.5 Dimensionality Reduction

Dimensionality reduction is used here to visualize the results of feature extraction in two dimensions. Seven different approaches to dimensionality reduction are implemented and compared in this work: six different manifold learning algorithms as well as PCA.

3.5.1 Manifold Learning

Manifold learning is an approach to dimensionality reduction that works on the assumption that data in high dimensional spaces can be embedded on a lower dimensional, non-linear, manifold within that space. If this lower dimensional manifold is two or three dimensions, then it can be visualized. One implementation of manifold learning that has shown promising results and has been used in previous work for visualization of drum samples [11] is t-distributed Stochastic Neighbour Embedding (t-SNE) [27]. Because t-SNE has a reduced tendency to crowd data points near the centre of the map, visualizations are considered to be significantly improved compared to its predecessor Stochastic Neighbor Embedding [28]. In order to more thoroughly investigate the efficacy of dimensionality reduction for visualization of drum audio samples, several other manifold learning algorithms have been included for testing. In total, six different manifold methods are compared here: t-SNE with random initialization, t-SNE with PCA initialization, Isomap [29], Locally-Linear Embedding (LLE) [30], Multi-dimensional Scaling (MDS) [31], and Spectral Embedding [32].

3.5.2 Principal Component Analysis

Principal Component Analysis (PCA) has been used in previous work as a dimensionality reduction technique for visualization of audio [13], and is included here for comparison. Because PCA is linear, more complex data structures existing in higher dimensions that another manifold learning technique might be able to catch can potentially be lost by PCA. Dimensionality reduction from 133-dimensions was computed for all seven methods and all time segmentations. Prior to dimensionality reduction, the feature variables were standardized to have a mean equal to zero and unit standard deviation.

4. ANALYSIS

4.1 Feature Analysis using PCA

In this section we perform an analysis of the time segmentation methodology using PCA. In addition to being used as a dimensionality reduction technique, PCA is used in exploratory data analysis and has been used as such in related IMP work by Wilson et al. [33]. In their work, Wilson et al. looked at 1501 different mixes of 10 different songs and used feature extraction and PCA to characterize these mixes and to determine to most relevant features in terms of variance. Similarly, we use feature extraction and PCA here to characterize kick and snare samples and to explore the effects of time segmentation in terms of variance. Our goal is to create a two-dimension representation that accurately captures the similarities and variations between drum sounds. Although PCA explained variance might not be relevant for classification, it does provide an indication that the identified features are more appropriate for capturing the similarities and variations between drum sounds across the entire data-set.

The principal components that result from PCA are a set of new axes that maximize the variance of the dataset such that the first axis contains the most variance. Using this, the results from feature extraction can be analyzed and the most relevant features for characterizing kick and snare samples can be determined in terms of variance. PCA was run independently on kick and snare drum samples for each time segmentation method. For each of these analyses the null hypothesis was rejected using Bartlett's test of sphericity, calculated using the NumPy package [34]. All feature variables were standardized prior to PCA to have a mean equal to zero and unit standard deviation.

Results of PCA give insight into how the time segmentation effects variance and which features are most useful for characterizing kick and snare drum samples. Variance is maximized in the first two dimensions when using a 100ms window starting at 20% of the attack for kick sounds, and a 250ms window starting at 90% of the attack for snare sounds. The first two dimensions explain 31.74% and 32.79% of the variance for kick and snare drums respectively. The main contributing features for the first dimension of kick drums after PCA are the mean and standard deviation of the HFC, and the high spectral energyband. The second MFCC band and the mean and standard deviation of the middle low spectral eneryband are main contributing features to the second dimension of kick drums. For snare drums, spectral energy (SE), and the 18th and 19th bark bands contribute highly to the first dimension and the standard deviation of bark spread, the standard deviation of the zero crossing rate, and MFCC band 5 are main contributors to the second dimension. Table 1 summarizes results for kick and snare drum analyses and notes the features that contribute highly to the associated dimensions for the time segmentation methods that retain the most variance. Dimensions three and four are also included to show how time segmentation effects the higher dimensions resulting from PCA.

5. EXPERIMENTAL RESULTS

5.1 Audio Classification

Audio classification is used here to evaluate and compare the effect of the time segmentation choices. The classification tasks performed are sample type, drum machine, and

Table 1: Principle Component Analysis: Variance Ratios

Length	Start	Kick					Snare				
		Dim 1	**Dim 2**	**Dim 3**	**Dim 4**	**1+2**	**Dim 1**	**Dim 2**	**Dim 3**	**Dim 4**	**1+2**
25ms	20%	16.17%	13.66%	8.87%	7.49%	29.83%	17.11%	13.85%[5]	9.51%	5.48%	30.97%
25ms	50%	17.75%	11.88%	9.50%	7.63%	29.63%	18.08%	13.73%	9.39%	5.37%	31.81%
25ms	90%	16.49%	11.19%	9.82%	7.33%	27.68%	19.38%	12.81%	8.79%	5.36%	32.18%
100ms	20%	17.58%	14.16%[2]	9.38%	7.76%	31.74%[3]	20.16%	11.03%	9.85%	5.93%	31.19%
100ms	50%	17.10%	13.41%	9.31%	8.19%	30.50%	20.75%	10.75%	9.77%	5.95%	31.32%
100ms	90%	16.51%	12.05%	10.38%	8.90%	28.56%	21.22%	10.5%	9.37%	6.26%	31.37%
250ms	20%	17.28%	14.07%	8.75%	8.16%	31.35%	21.22%	10.73%	9.22%	6.52%	31.95%
250ms	50%	16.87%	13.48%	8.86%	8.22%	30.35%	21.70%	10.54%	9.08%	6.52%	32.24%
250ms	90%	16.33%	12.52%	9.51%	8.60%	28.85%	22.75%[4]	10.04%	8.89%	6.63%	32.79%[6]
500ms	20%	17.34%	13.38%	8.74%	7.82%	30.71%	21.43%	10.19%	9.15%	6.94%	31.62%
500ms	50%	16.84%	12.99%	8.71%	8.13%	29.84%	21.86%	10.01%	9.03%	6.97%	31.87%
500ms	90%	16.01%	12.41%	9.36%	8.52%	28.42%	22.70%	9.69%	8.71%	7.01%	32.39%
Mixed	-	16.17%	13.66%	8.87%	7.49%	29.83%	18.04%	11.00%	10.04%	6.55%	29.04%
Full	0%	18.08%[1]	13.46%	9.00%	7.37%	31.47%	21.01%	10.38%	9.15%	7.08%	31.54%

Table shows how much each dimension contributes in terms of variance after principal component analysis. The highlighted cells show the time segmentation combination that contains the most variance for that column. The features that contributed the most to the variance for each of those highlighted cells is shown below:

Main contributing features:

[1] **Dim 1:** HFC, HFC Std Dev, Mid-High Spectral Energyband

[2] **Dim 2:** MFCC Band 2, Mid-Low Spectral Energyband Std Dev, Mid Low Spectral Energyband

[3] **Dim 1:** HFC, HFC Std Dev, High Spectral Energyband; **Dim 2:** MFCC Band 2, Mid-Low Spectral Energyband Std Dev, Mid Low Spectral Energyband

[4] **Dim 1:** Spectral Energy, Bark Band 18 and 19

[5] **Dim 2:** Spectral Decrease, Spectral Decrease Std Dev, Spectral RMS

[6] **Dim 1**:Spectral Energy, Bark Band 18 and 19 **Dim 2**: Bark Spread Std Dev, Zero Crossing Rate Std Dev, MFCC Band 5

manufacturer classification. Three different classification algorithms implemented in Scikit-learn [35] are used: Support Vector Machine, Perceptron, and Random Forest. 10-fold cross-validation was used for each classification task and accuracy scores are calculated as an average between the three algorithms. The ZeroR classifier, which simply predicts the majority class, was used to determine the baseline accuracy for each task. Classification was run on all time segmentation choices, as well as on the full sample duration and the mixed time segmentation method, and all training data was shuffled prior to training.

5.1.1 Sample Type Classification

Sample type classification seeks to distinguish between kick and snare samples. All of the samples from the dataset were used and the baseline accuracy score was calculated to be 52.13%. The highest accuracy for kick and snare classification was 97.8% and was achieved by three time segmentation methods: 250ms positioned at 50% the attack, 500ms positioned at 50% the attck, and mixed time segmentations. Full results are shown under the Sample Type column in Table 2.

Table 2: Accuracy scores in percentages for classification tests trained using results from feature extraction and time segmentation. Mean and standard deviation of accuracy across validation folds are shown. The shaded cells show the highest performing time segmentation for each task (multiple cells shown in case of a tie).

Length	**Start**	**Sample Type**	**Drum Machine**		**Manufacturer**	
			Kick	**Snare**	**Kick**	**Snare**
25ms	20%	95.5 ± 1.8	87.5 ± 7.1	75.7 ± 5.1	50.4 ± 12.2	47.7 ± 10.6
25ms	50%	96.1 ± 1.7	86.5 ± 5.3	73.3 ± 5.6	51.1 ± 12.7	46.7 ± 9.6
25ms	90%	96.4 ± 2.1	82.6 ± 6.1	70.5 ± 6.2	47.5 ± 11.1	48.2 ± 9.4
100ms	20%	97.6 ± 1.3	93.4 ± 5.1	78.3 ± 4.8	51.8 ± 13.7	50.6 ± 10.5
100ms	50%	97.7 ± 1.6	90.5 ± 5.3	76.3 ± 5.0	52.2 ± 12.3	49.7 ± 11.3
100ms	90%	97.5 ± 2.4	90.0 ± 5.4	75.1 ± 5.4	48.7 ± 12.2	51.0 ± 10.6
250ms	20%	97.6 ± 1.6	92.7 ± 4.2	76.8 ± 5.7	52.1 ± 12.6	51.3 ± 9.5
250ms	50%	97.8 ± 1.2	92.2 ± 4.0	77.6 ± 4.6	53.3 ± 12.8	49.9 ± 9.6
250ms	90%	97.7 ± 1.4	87.4 ± 7.1	73.5 ± 4.5	50.9 ± 10.7	51.3 ± 9.9
500ms	20%	97.5 ± 1.8	93.5 ± 3.4	80.7 ± 4.4	52.3 ± 12.2	52.3 ± 8.8
500ms	50%	97.8 ± 1.3	92.2 ± 3.7	80.2 ± 4.7	53.2 ± 11.6	51.4 ± 9.8
500ms	90%	97.7 ± 1.5	88.9 ± 5.7	79.6 ± 5.5	50.0 ± 12.3	54.3 ± 7.9
Mixed	-	97.8 ± 1.3	94.0 ± 4.4	81.1 ± 4.2	53.0 ± 13.6	51.8 ± 8.8
Full	0	97.5 ± 1.6	93.5 ± 4.7	80.9 ± 4.9	54.8 ± 13.6	55.1 ± 7.9

5.1.2 Drum Machine Classification

Drum machine and manufacturer classification tasks performed within each sample class based on the classification study on unpitched percussion sounds by Herrera et al. [15] are used to evaluate the ability of this analysis technique to characterize percussion samples of the same type.

For drum machine classification, machines were selected for kicks and snares separately such that each drum machine would have at least 50 samples for each type. Six distinct classes were used for kick drums which resulted in 464 kick samples in total and a baseline accuracy of 22.20%. The drum machines selected for kick drum classification were the Alesis DM5, Alesis SR-16, Roland SH-09, Roland TR-808, Roland TR-909, and the Yamaha RM50. Nine distinct classes were used for snare drums which resulted in 726 snare samples and a baseline accuracy of 14.88%. Drum machines used for the snare drum classification were the Alesis DM5, Alesis SR-16, Boss DR-660, Roland System-100, Roland TR-808, Roland TR-909, Yamaha CS6, Yamaha RY30, and the Yamaha RM50. Classification performed best using the mixed time segmentation scheme for both kicks and snares, reporting 94% and 81.1% accuracy respectively. Full results are shown under the Drum Machine column shown on Table 2.

Confusion matrices give further insight into the performance of each classification task by providing a breakdown of how each sample was categorized. In the specific case, matrices are produced by averaging the classification results from the three algorithms used. Confusion matrices produced from the results of the mixed time segmentation test for both kicks and snares are shown in tables 5a and 6a respectively. These results show that for kick drums, the Yamaha RM50 was easiest to classify with 102 of the 103 samples included being labelled correctly. For snare drums, sounds produced by the System-100 were the easiest to classify with 71 out of 72 samples being labelled correctly. Interestingly, the Yamaha RM50 snare sounds were more challenging to classify with only 26.33 out of 67 samples being classified correctly. Both kicks and snares from the iconic Roland TR-808 drum machine were guessed accurately with 62.67 out of 67 kick samples labelled correctly and 58.33 out of 63 snare samples labelled correctly.

5.1.3 Manufacturer Classification

Manufacturers were selected such that each manufacturer was represented by at least 100 samples of each type. The same six manufactures were used for the kicks and snares and are Alesis, Boss, E-MU, Korg, Roland, and Yamaha. For kick drums a total of 1328 samples were included reporting a baseline accuracy score of 39.16%. For snare drums a total of 1556 samples were used reporting a baseline accuracy of 33.1%. Results show that manufacturer classification was a more difficult task than the previous two tasks; kick drum classification reported 54.8% accuracy using the full sample length, and snares reported 55.1% accuracy using the full sample length. Full results are shown under the Manufacturer column on Table 2.

5.2 Evaluation of Dimensionality Reduction

Classification is also used to evaluate the effectiveness of the seven methods of dimensionality reduction implemented. The drum machine classification test is repeated, however, instead of using the full 133-dimension feature vectors to train classifiers, only the first two dimensions resulting from dimensionality reduction are used as training data. Using the scores achieved during the full dimension tests as a baseline, the amount of information lost during dimensionality reduction, as well as the overall effectiveness of each method, can be quantified. In addition, each of the 12 time segmentation methods, as well as the full sample length and the mixed time segmentation method, are included to further evaluate the effectiveness of time segmentation. In total, 98 classification tests were run for each kick and snare sample types. For both kick and snare drums, classification accuracy scores were maximized when using t-SNE, with both random initialization and PCA initialization producing similar results. The kick drum test produced an accuracy score of 76.3% using mixed time segmentation for both t-SNE initialization strategies, compared to a score of 94% when using the full dimension feature vector and a mixed time segmentation. The snare drum test produced an accuracy score of 58.6% using a time segment of 500ms positioned at 20% the attack and t-SNE with random intialization, compared to a score of 71.21% when using the full dimension feature vector and a mixed time segmentation. Full results for these tests are shown on Tables 3 and 4 for kicks and snares respectively.

Confusion matrices were produced for both kick and snare classification tests using t-SNE with random initialization and mixed time segmentations. Comparing the confusions matrices from before and after dimensionality reduction gives further insight into how reducing to two-dimensions affects classifier performance. For kick drums, 89.67 out of 103 Yamaha RM50 samples were labelled correctly after dimensionality reduction compared to 102 out of 103 before. For snare drums, only 21.33 out of 67 Yamaha RM50 samples were labelled correctly after dimensionality reduction compared 26.33 out of 67 before. 62.67 out

Table 3: Tables 3 & 4 show accuracy scores in percentages for drum machine classification task trained using the dimensionality reduced data from manifold learning for kicks and snares respectively. Mean and standard deviation across validation folds is shown for each combination of manifold learning and time segmentation method. The shaded cell shows the highest overall score (multiple cells shown in case of a tie).

Length	Start	PCA	Isomap	LLE	MDS	Spectral	TSNE	TSNE PCA
25ms	20%	46.6 ± 8.0	46.3 ± 7.3	54.2 ± 6.9	46.6 ± 8.6	51.6 ± 8.5	65.1 ± 5.7	64.1 ± 7.5
25ms	50%	46.5 ± 7.7	43.0 ± 6.3	48.7 ± 8.8	42.9 ± 10.1	47.1 ± 8.4	64.7 ± 6.2	62.7 ± 8.2
25ms	90%	45.7 ± 9.5	39.4 ± 5.6	48.3 ± 6.7	53.7 ± 7.0	48.2 ± 8.6	55.8 ± 6.3	57.0 ± 5.9
100ms	20%	55.4 ± 8.0	49.5 ± 6.7	63.1 ± 9.2	62.8 ± 7.7	58.7 ± 7.3	73.9 ± 9.7	71.4 ± 7.1
100ms	50%	52.0 ± 6.4	39.4 ± 6.6	62.4 ± 7.9	63.2 ± 6.8	60.0 ± 10.5	75.6 ± 8.4	74.5 ± 8.4
100ms	90%	61.8 ± 8.0	56.5 ± 7.1	52.6 ± 8.2	62.6 ± 6.2	58.5 ± 6.6	70.2 ± 7.0	71.8 ± 7.4
250ms	20%	53.8 ± 8.6	55.4 ± 6.6	60.4 ± 9.5	57.5 ± 8.3	60.5 ± 8.8	70.3 ± 7.9	69.8 ± 8.3
250ms	50%	54.5 ± 9.6	42.2 ± 5.8	55.1 ± 8.5	56.0 ± 7.9	60.9 ± 8.5	70.5 ± 9.5	70.0 ± 9.2
250ms	90%	60.4 ± 6.5	35.4 ± 6.8	56.1 ± 8.6	54.8 ± 9.4	57.8 ± 5.9	72.2 ± 6.6	71.8 ± 5.9
500ms	20%	59.3 ± 8.2	44.3 ± 5.7	59.5 ± 8.5	57.0 ± 7.6	63.1 ± 9.2	71.9 ± 9.6	69.6 ± 8.6
500ms	50%	61.3 ± 8.2	48.8 ± 6.8	56.7 ± 9.7	55.6 ± 8.5	59.7 ± 8.0	71.4 ± 6.5	71.4 ± 7.9
500ms	90%	59.1 ± 6.6	51.7 ± 6.3	52.6 ± 8.7	55.0 ± 5.6	59.1 ± 8.4	70.5 ± 8.3	71.4 ± 8.2
Mixed	-	59.1 ± 9.8	52.2 ± 9.5	60.7 ± 8.5	62.7 ± 9.6	59.9 ± 7.6	76.3 ± 8.9	76.3 ± 9.0
Full	0%	58.9 ± 7.2	48.1 ± 7.3	60.9 ± 10.5	59.8 ± 9.3	65.6 ± 8.8	72.5 ± 9.1	72.0 ± 7.6

Table 4: Accuracy scores from reduced dimension drum machine classification task for snare drums.

Length	Start	PCA	Isomap	LLE	MDS	Spectral	TSNE	TSNE PCA
25ms	20%	41.1 ± 5.3	28.5 ± 4.3	35.9 ± 5.7	37.1 ± 5.2	33.8 ± 6.0	51.4 ± 4.9	50.1 ± 5.0
25ms	50%	38.2 ± 3.8	32.5 ± 4.6	34.2 ± 5.1	37.2 ± 6.1	34.4 ± 5.2	48.9 ± 6.2	49.3 ± 6.2
25ms	90%	35.9 ± 4.8	23.5 ± 4.3	36.1 ± 5.8	33.4 ± 5.2	32.0 ± 4.1	46.7 ± 6.7	47.7 ± 5.5
100ms	20%	40.7 ± 5.2	32.6 ± 5.8	42.5 ± 5.8	36.8 ± 5.0	37.7 ± 5.0	56.1 ± 4.9	56.6 ± 5.5
100ms	50%	42.5 ± 6.1	33.7 ± 5.8	38.8 ± 5.4	36.0 ± 4.0	37.9 ± 5.8	57.4 ± 5.3	58.5 ± 6.1
100ms	100%	42.2 ± 5.9	31.9 ± 4.7	39.8 ± 5.1	33.9 ± 4.5	35.1 ± 5.9	52.8 ± 5.7	53.7 ± 5.6
250ms	20%	39.2 ± 5.3	35.3 ± 5.1	33.0 ± 5.2	35.8 ± 5.5	33.8 ± 5.3	58.0 ± 6.4	57.5 ± 5.6
250ms	50%	38.6 ± 4.9	29.1 ± 4.3	35.8 ± 5.6	35.3 ± 5.1	33.9 ± 5.3	54.5 ± 6.1	55.8 ± 6.3
250ms	90%	37.5 ± 5.7	35.3 ± 4.3	34.7 ± 5.6	37.3 ± 5.2	33.2 ± 4.9	52.8 ± 6.9	53.1 ± 6.7
500ms	20%	40.7 ± 4.9	33.5 ± 7.3	38.1 ± 6.8	34.9 ± 5.3	34.5 ± 5.1	58.6 ± 7.1	57.1 ± 6.2
500ms	50%	39.8 ± 4.9	33.6 ± 4.4	36.9 ± 6.0	36.7 ± 5.7	33.0 ± 4.6	56.3 ± 6.5	58.0 ± 5.8
500ms	90%	38.6 ± 6.2	32.5 ± 4.6	36.9 ± 3.9	33.9 ± 6.5	34.1 ± 4.5	53.4 ± 7.0	54.3 ± 5.5
Mixed	-	44.9 ± 5.0	33.9 ± 7.2	41.2 ± 7.3	35.8 ± 6.1	36.4 ± 5.8	56.4 ± 5.4	57.5 ± 6.1
Full	0%	41.7 ± 4.5	34.8 ± 5.3	38.8 ± 5.3	36.2 ± 5.0	35.2 ± 4.3	56.2 ± 5.7	56.5 ± 6.9

of 72 System-100 snare drum samples were labelled correctly after dimensionality reduction compared to 71 out of 72 before. Full confusion matrix results for these tests are shown for kicks and snares in Tables 5b and 6b respectively.

Audio samples used in the drum machine classification tests have been plotted using the results from dimensionality reduction. The top method and time segmentation combination are shown for both kicks and snares and a plot of the same time segmentation scheme using PCA is also included for comparison. Kick drums plotted using PCA and the mixed time segmentation scheme are shown in Figure 3a, kick drums using t-SNE with random initialization and the mixed time segmentation scheme are shown in Figure 3b, snare drums

Table 5: Kick drum machine classification confusion matrices. The dimensionality reduced results for both kicks and snares were produced by running classification on the results of t-SNE with random initialization using the mixed time segmentation method. Each row shows how samples from that drum machine were classified and highlighted cells along the diagonal are the number of correctly classified samples. Note that fractions are caused by averaging between results from the three classification methods used.

(a) Kick Drums - Full Dimension

	DM5	SR-16	SH-09	TR-808	TR-909	RM50
DM5	84.67	2.33	1.33	2.67	3.00	1.00
SR-16	4.33	57.33	0.00	0.00	0.33	0.00
SH-09	0.00	0.00	80.33	5.67	0.00	0.00
TR-808	3.00	0.00	1.00	62.67	0.33	0.00
TR-909	2.33	0.00	0.33	2.00	46.33	0.00
RM50	0.00	1.00	0.00	0.00	0.00	102.00

(b) Kick Drums - Dimensionality Reduced

	DM5	SR-16	SH-09	TR-808	TR-909	RM50
DM5	56.67	18.33	4.33	6.33	4.67	4.67
SR-16	13.67	45.00	0.67	2.00	0.00	0.67
SH-09	0.33	5.00	61.67	18.67	0.33	0.00
TR-808	2.33	0.00	5.00	59.00	0.67	0.00
TR-909	1.67	4.67	10.67	2.67	31.33	0.00
RM50	0.00	13.33	0.00	0.00	0.00	89.67

Table 6: Snare drum machine classification confusion matrices. The dimensionality reduced results for both kicks and snares were produced by running classification on the results of t-SNE with random initialization using the mixed time segmentation method. Each row shows how samples from that drum machine were classified and highlighted cells along the diagonal are the number of correctly classified samples. Note that fractions are caused by averaging between results from the three classification methods used.

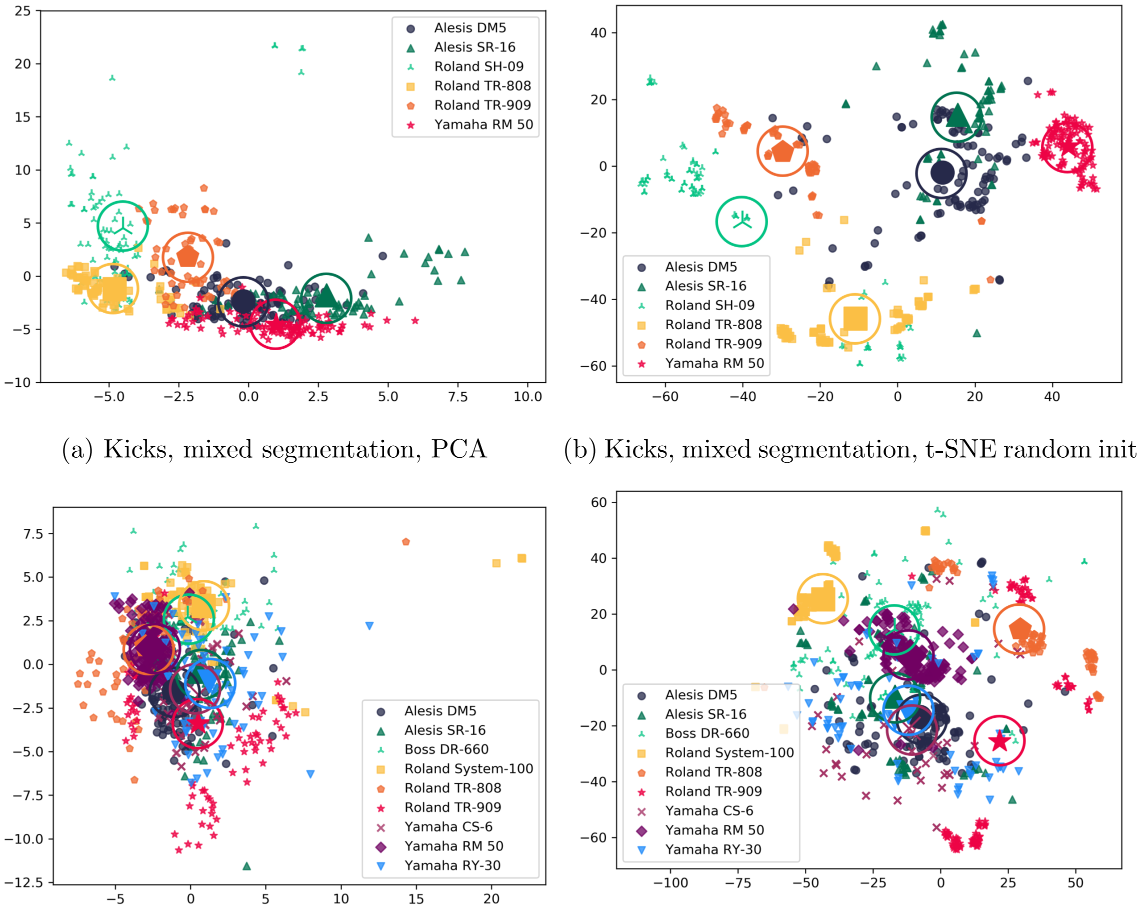
(a) Snare Drums - Full Dimension

	DM5	SR-16	DR-660	Sys-100	TR-808	TR-909	CS6	RY30	RM50
DM5	87.67	4.00	3.33	0.33	0.33	0.33	3.00	3.67	5.33
SR-16	1.33	47.67	2.33	0.00	0.67	0.67	2.33	9.67	6.33
DR-660	0.67	5.33	60.67	1.00	1.33	0.67	0.67	4.67	2.00
Sys-100	0.33	0.00	0.00	72.00	0.00	0.00	0.00	0.67	0.00
TR-808	0.00	1.33	0.67	0.00	58.33	0.00	0.67	0.00	2.00
TR-909	0.00	0.33	1.33	0.00	0.00	87.67	0.00	0.00	0.67
CS6	3.67	6.00	1.33	0.00	2.67	0.67	51.00	0.00	3.67
RY30	5.33	2.00	3.33	1.00	0.00	0.00	1.00	95.33	0.00
RM50	6.33	14.33	2.67	0.00	1.67	2.00	8.67	5.00	26.33

(b) Snare Drums - Dimensionality Reduced

	DM5	SR-16	DR-660	Sys-100	TR-808	TR-909	CS6	RY30	RM50
DM5	64.67	11.67	6.67	0.00	3.00	5.00	7.00	6.00	4.00
SR-16	15.67	24.00	3.67	3.67	2.33	2.33	4.00	10.33	5.00
DR-660	8.67	9.00	30.67	5.00	4.67	1.67	1.00	13.00	3.33
Sys-100	0.67	5.00	0.67	62.67	3.00	0.00	0.33	0.67	0.00
TR-808	2.33	1.67	0.00	1.00	49.33	4.67	0.33	3.67	0.00
TR-909	6.00	3.33	0.00	0.00	11.33	64.67	3.33	1.33	0.00
CS6	17.33	6.00	1.33	0.00	5.67	4.33	30.33	2.00	2.00
RY30	13.67	11.67	6.33	1.00	0.33	1.00	1.33	71.67	1.00
RM50	14.33	8.67	4.67	1.00	0.33	4.33	5.33	7.00	21.33

plotted using PCA and mixed time segmentations are shown in Figure 3c, and snare drums plotted using t-SNE with random initialization and mixed time segmentations are shown in Figure 3d. From a visual analysis, the t-SNE algorithms appear to improve separation between samples clustered by drum machines. These findings are consistent with the claim that t-SNE has a reduced tendency to crowd data points in the centre of the map, and in the specific case, this has also translated in an improvement in classification scores compared to other manifold learning techniques as well as PCA.



(a) Kicks, mixed segmentation, PCA

(b) Kicks, mixed segmentation, t-SNE random init

(c) Snares, mixed segmentation, PCA

(d) Snares, mixed segmentation, t-SNE random init

Figure 3: Scatter plots of samples used in the drum machine classification tasks. Samples are visualized for kicks and snares using PCA as well as t-SNE with random initialization. The mixed time segmentation scheme is shown for all plots. The centroid for each drum machine is marked with a larger symbol and is circled. Note that dimensions are left unlabelled; after dimension reduction the 'x' and 'y' axis correspond to the 1st and 2nd dimensions produced by the dimensionality reduction algorithms.

6. Subjective Testing

6.1 Design

A subjective listening test was undertaken to ascertain if the methods we use to generate the intuitive browsing system is perceptually relevant to users. In the test, participants were asked to rank the similarity of kick or snare drum sounds to a reference sound. This question was designed to put listeners into an integrative state, such that they treat the sounds as a whole, and don't focus upon individual perceptual features [36].

6.2 Participants

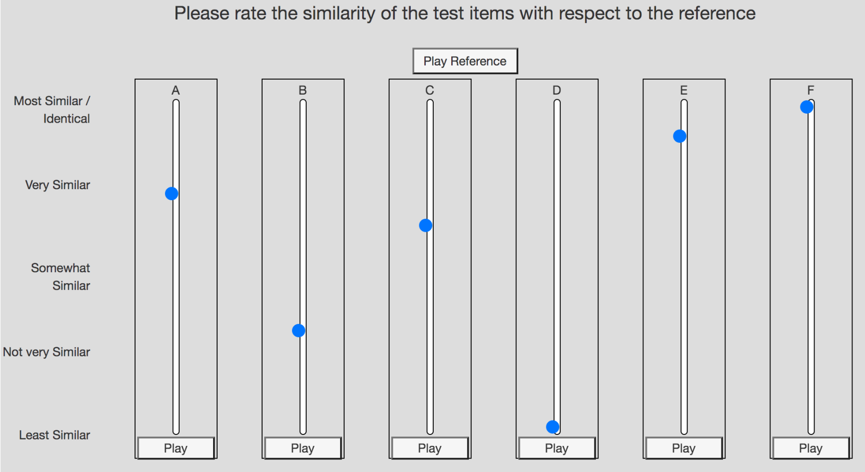
Participants were recruited for the listening test using professional and research networks, and online forums focussing on music production for genres where the use of drum samples is prevalent. The listening test was approved by the human research ethics board at the University of Victoria (ID 20-0101), with participants providing informed consent via an online form as part of the listening test procedure. A total of 46 participants completed the test, 6 were female, 37 were male, and 3 were non-binary or chose to not provide gender information. The age range was 21 to 60, with a median age of 30. The average reported experience of participants using kick and snare drum samples in music productions was 7.5 years, with 39 participants reporting having one year or more of experience working with drum samples. 33 participants reported that they currently produced, mixed, or remixed music that uses drum samples.

6.3 Materials

A set of 16 reference sounds (8 kick and 8 snare) were selected using the output from the feature extraction process described in section 3.2, with no time segmentation applied. In order to ensure that these samples were representative of the variety of drum samples found in the larger collection (4230), eight equally sized groups were created for both kicks and snares using distance measures from the centroid of the entire data set. Samples with a distance in the top fifth percentile were considered outliers and were not considered for reference selection. One sample from each of these eight groups was then randomly selected to be a reference. Once 16 references had been selected, five stimuli were selected such that each stimuli was at ever increasing percentile distances from the reference up to the sample at the 90th percentile distance. This provided the materials for a total of 16 individual listening tests, eight using kick drum samples and eight using snare drum samples, and a total of 80 unique sample pair similarity ratings.

6.4 Apparatus

A MUSHRA style [37] listening test was used for this study, created using the Web-Audio Evaluation Tool [38]. The MUSHRA test format was selected because it allows for the comparison of a collection of sounds, which can be seen as representative of a visual browsing system, and has been shown as effective in a similar study evaluating the similarity between sounds [39]. Participants were presented with the apparatus shown in figure 4, and asked to use the sliders to rank the similarity of the samples to the reference sound. The vertical



Please rate the similarity of the test items with respect to the reference

Figure 4: Interface of the listening test apparatus, which was administered online and built using the Web Audio Evaluation Tool. Test was a MUSHRA style listening test with five stimuli and one hidden reference. Participants were asked to rate each sample to the reference based on sound similarity

continuous, unnumbered slider was labelled with ”identical” for the maximum and ”least similar” for the minimum values. These labels were chosen to encourage participants to use the end points of the scale, to prevent a clustering of responses in the middle of the scale [36], and as a means of reliably reporting when they were able to identify the hidden reference, which was included for all tests.

6.5 Procedure

All listening tests were conducted remotely using a webpage [6]. Participants were allowed to set their desired playback level, and use equipment and environments that they were familiar and comfortable with. The test contained a total of 16 pages, each with one reference, and a random selection of 8 pages was presented to each user to reduce the amount of time per test. Listeners were instructed to use the interface to rank the similarity of test items compared to the reference, with the system ensuring each slider had been moved and each test item had been played before allowing the participant to continue to the next page of the test. All ratings were recorded on a scale from 0 to 1, with 0 being ’least similar’ and 1 being ’most similar / identical’.

6.6 Results

Each test page received a minimum of 21 responses and an average of 23 responses per page. The average time taken to complete the listening test was 8.4 minutes. To determine if listeners consistently ranked drum samples as similar, the results were analyzed using weighted Spearmen correlation coefficients to determine Kendall’s coefficient of concordance [40, 41, 42]. This method was necessary because of the experimental design, which

[6] `https://listeningtest.uvic.ca`

resulted in a non-uniform distribution of page responses. Results of the analysis show that participants rated drum samples significantly similar (0.39, p∼0). The same analysis was used to evaluate the effect of the self-reported experience using drum samples by listening test subjects, and interestingly listeners with no experience rated drum samples more similarly (0.42, p∼0) than listeners with experience using drum samples (0.40, p∼0).

Correlation coefficients between the average sample similarity ranking from the listening test and the similarity measure taken from the audio feature extraction was used to investigate how well the full dimensional results correlate with the subjective similarity rankings, and how this compared against the time-segmentation and the different dimensionality reduction methods. Computational similarity was calculated as the negative of the Euclidian distance between samples. The correlation coefficient between the average ranking for a sample and the similarity measure from the feature extraction with no time segmentation or dimensionality reduction was 0.8717. Correlation was maximized on the full dimension data using the mixed time segmentation method (0.8935). The dimensionality reduction method that performed the best was MDS using a time segmentation with a sample length of 100ms starting at 20% of the sample attack, which resulted in a correlation coefficient of 0.8329. The method that performed the worst was locally linear, with an average correlation coefficient of 0.32. The heatmap in Figure 5 shows the correlation coefficients for all dimensionality reduction and time segmentation methods.

7. Software Implementation

The methods described in this paper have been integrated into a plugin developed by the authors using the JUCE framework[7]. The goal of the plugin is to provide music producers with a more efficient and intuitive method to search and audition drum samples within digital audio workstations. Samples are automatically organized based upon sound similarity and visually represented using a 2D grid, which can then be mapped to a hardware controller, such as the Ableton Push controller. For a complete overview of the software implementation please see [6]. Source code and documentation is available at `https://github.com/jorshi/sieve`.

8. Reproducibility

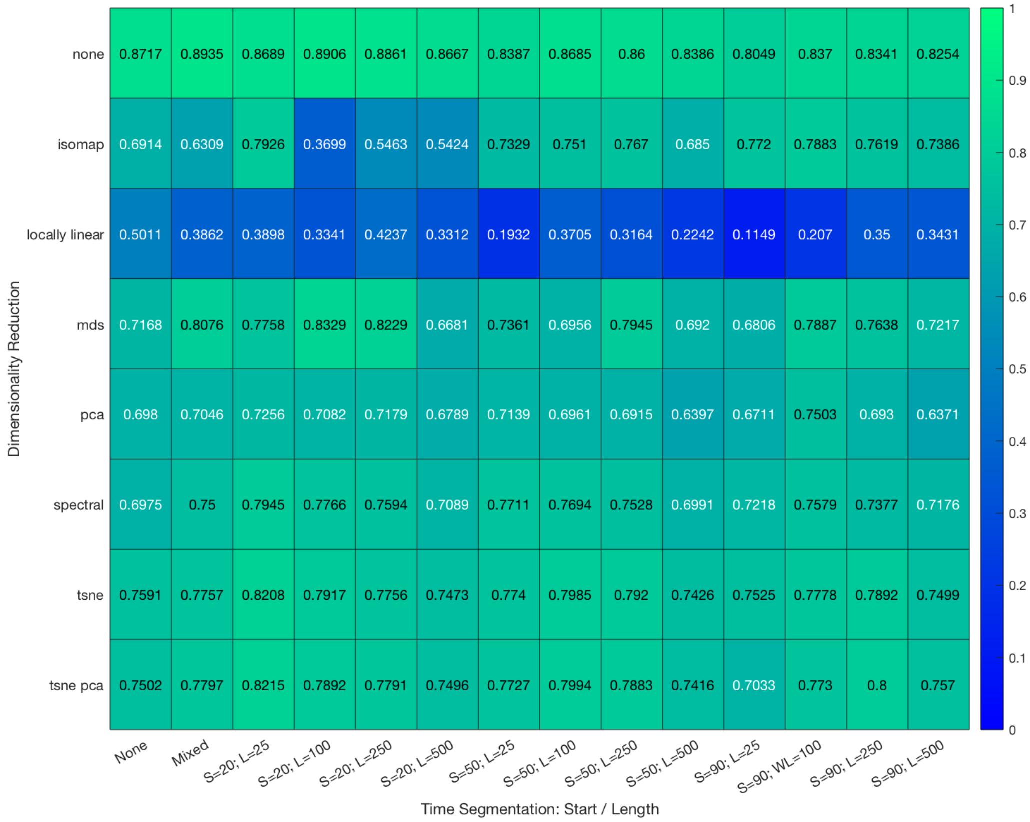
The authors welcome any feedback and contributions on the GitHub page [8] in accordance with the recommendations for open access and reproducibility in signal processing research presented by Vandewalle et al. [43]. The dataset is also available upon request.

9. CONCLUSION

In this paper we presented a methodology for computationally characterizing and organizing kick and snare drum samples in two-dimensions based on sound similarity. The goal of this methodology is to support the development of intuitive audio sample browsing systems that improve upon the current approach of browsing through lists or searching using text

[7] `https://www.juce.com/`

[8] `https://github.com/jorshi/sample_analysis`



	None	Mixed	S=20; L=25	S=20; L=100	S=20; L=250	S=20; L=500	S=50; L=25	S=50; L=100	S=50; L=250	S=50; L=500	S=90; L=25	S=90; WL=100	S=90; L=250	S=90; L=500
none	0.8717	0.8935	0.8689	0.8906	0.8861	0.8667	0.8387	0.8685	0.86	0.8386	0.8049	0.837	0.8341	0.8254
isomap	0.6914	0.6309	0.7926	0.3699	0.5463	0.5424	0.7329	0.751	0.767	0.685	0.772	0.7883	0.7619	0.7386
locally linear	0.5011	0.3862	0.3898	0.3341	0.4237	0.3312	0.1932	0.3705	0.3164	0.2242	0.1149	0.207	0.35	0.3431
mds	0.7168	0.8076	0.7758	0.8329	0.8229	0.6681	0.7361	0.6956	0.7945	0.692	0.6806	0.7887	0.7638	0.7217
pca	0.698	0.7046	0.7256	0.7082	0.7179	0.6789	0.7139	0.6961	0.6915	0.6397	0.6711	0.7503	0.693	0.6371
spectral	0.6975	0.75	0.7945	0.7766	0.7594	0.7089	0.7711	0.7694	0.7528	0.6991	0.7218	0.7579	0.7377	0.7176
tsne	0.7591	0.7757	0.8208	0.7917	0.7756	0.7473	0.774	0.7985	0.792	0.7426	0.7525	0.7778	0.7892	0.7499
tsne pca	0.7502	0.7797	0.8215	0.7892	0.7791	0.7496	0.7727	0.7994	0.7883	0.7416	0.7033	0.773	0.8	0.757

Figure 5: Heatmap showing the correlation coefficients between average similarity ranking from the listening test and computational similarity measurement using a specific combination of dimensionality reduction and time segmentation. The top row shows values for no dimensionality reduction where similarity ranking was computed on the full-dimension feature set.

or semantic tags. In this work, the use of time segmentation was explored, which isolates a temporal subset of an audio sample prior to audio feature extraction. Several different segmentations comprised of various lengths and start position offsets were compared. Time segmentation was shown to improve classification scores and produce sample similarity ratings that correlate more highly with subjective rankings when compared to results using the full sample duration. A mixed time segmentation method was introduced that selects a segmentation for each audio feature independently to maximize variance. Objective evaluation using a drum machine classification task showed that using a mixed time segmentation approach resulted in the highest accuracy when trained on the full dimension dataset. Using mixed time segmentations also lead to the highest classification scores for kick drums when trained using the dimensionality reduced data.

Manifold learning was explored to reduce the high dimensional results from audio feature extraction down to two dimensions for visualization, and several different manifold learning algorithms were compared. Audio classification results showed that the t-SNE manifold learning algorithm outperformed all other dimensionality reduction techniques compared in this study. Qualitative evaluation using visual plots of the audio samples after dimensionality reduction confirmed the classification results, where centroid values for each drum machine are observed as being spaced further apart from each other, creating a clearer representation of the 'sonic layout' of the sample set.

A listening test was carried to generate ground truth similarity rankings for a set of kick and snare drum samples to evaluate the perceptual relevance of the presented methodology. Findings show that the mixed time segmentation approach with no dimensionality reduction produced drum sample similarity scores that correlate most highly with the subjective similarity rankings. The MDS algorithm with time segmentation produced an organization of samples in two-dimensions that correlated most highly the subjective rankings.

These results show that time segmentation is a beneficial step in the process of computationally characterizing and organizing drum samples in two-dimensions based on sound similarity. The use of mixed time segmentations produced promising results during the audio classification tasks and is an area for exploration in future related work. A limitation of the current implementation of the mixed time segmentation scheme is that it requires all features to be calculated for all time segmentations for a set of samples before time segmentations can be selected; the time and storage requirements of this approach might not be appropriate for a production ready application. Exploring techniques for improving the efficiency of this calculation is another area of future work.

Extending the scope of this work beyond kick and snare drum machine sounds would be a natural next step for building upon this methodology. Percussion instruments lend themselves well to our proposed method for calculating time segmentations due to their transient nature, however more work is required to explore the effectiveness with different percussion instruments as well as acoustic sounds. Extending this technique beyond percussion instruments would also be an interesting area of study. Exploring different approaches to using time segmentation could also lead to some valuable improvements. For example, instead of completely isolating a time segment prior to audio feature extraction, an alternative method could instead emphasize (or de-emphasize) the selected time segment in relation to the whole audio clip. This could be implemented in a way that would allow a user to selectively look for sounds and focus on certain temporal aspects (ex. a user looking for a kick drum sound with a specific attack character could indicate on the user interface that they are interested in that portion of the sound).

References

[1] K. Andersen and P. Knees, “Conversations with expert users in music retrieval and research challenges for creative mir.,” in *ISMIR*, pp. 122–128, August 2016.

[2] E. J. Humphrey, D. Turnbull, and T. Collins, “A brief review of creative mir,” *ISMIR Late-Breaking News and Demos*, November 2013.

[3] D. Moffat and M. B. Sandler, “Approaches in intelligent music production,” in *Arts*, vol. 8, p. 125, Multidisciplinary Digital Publishing Institute, September 2019.

[4] P. Knees and K. Andersen, “Searching for audio by sketching mental images of sound: a brave new idea for audio retrieval in creative music production,” in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 95–102, ACM, June 2016.

[5] J. Shier, K. McNally, and G. Tzanetakis, “Analysis of drum machine kick and snare sounds,” in *Audio Engineering Society Convention 143*, Audio Engineering Society, October 2017.

[6] J. Shier, K. McNally, and G. Tzanetakis, “Sieve: A plugin for the automatic classification and intelligent browsing of kick and snare samples,” in *3rd Workshop on Intelligent Music Production*, WIMP, September 2017.

[7] K. Andersen and F. Grote, “Giantsteps: Semi-structured conversations with musicians,” in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2295–2300, ACM, April 2015.

[8] M. Cooper, J. Foote, E. Pampalk, and G. Tzanetakis, “Visualization in audio-based music information retrieval,” *Computer Music Journal*, vol. 30, pp. 42–62, Summer 2006.

[9] M. Schedl, E. Gómez, J. Urbano, *et al.*, “Music information retrieval: Recent developments and applications,” *Foundations and Trends® in Information Retrieval*, vol. 8, pp. 127–261, September 2014.

[10] O. Fried, Z. Jin, R. Oda, and A. Finkelstein, “Audioquilt: 2D arrangements of audio samples using metric learning and kernelized sorting.,” in *New Interfaces for Musical Expression*, pp. 281–286, June 2014.

[11] C. Turquois, M. Hermant, D. Gómez-Marín, and S. Jordà, “Exploring the benefits of 2D visualizations for drum samples retrieval,” in *2016 ACM on Conference on Human Information Interaction and Retrieval*, pp. 329–332, ACM, March 2016.

[12] M. Cartwright, B. Pardo, and J. Reiss, “Mixploration: Rethinking the audio mixer interface,” in *The 19th international conference on Intelligent User Interfaces*, pp. 365–370, ACM, February 2014.

[13] G. Tzanetakis and P. Cook, “3D graphics tools for sound collections,” in *Conference on Digital Audio Effects, DAFX*, December 2000.

[14] P. Herrera, A. Yeterian, and F. Gouyon, “Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques,” in *Music and Artificial Intelligence*, pp. 69–80, Springer, September 2002.

[15] P. Herrera, A. Dehamel, and F. Gouyon, “Automatic labeling of unpitched percussion sounds,” in *Audio Engineering Society Convention 114*, Audio Engineering Society, March 2003.

[16] P. Herrera, V. Sandvold, and F. Gouyon, “Percussion-related semantic descriptors of music audio files,” in *Proc. of 25th International AES Conference*, Citeseer, June 2004.

[17] F. Tutzer *et al.*, "Drum rhythm retrieval based on rhythm and sound similarity," *Master's thesis, Departament of Information and Communication Technologies Universitat Pompeu Fabra, Barcelona*, September 2011.

[18] V. M. Souza, G. E. Batista, and N. E. Souza-Filho, "Automatic classification of drum sounds with indefinite pitch," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, pp. 1–8, IEEE, July 2015.

[19] P. Toiviainen, "Optimizing auditory images and distance metrics for self-organizing timbre maps," *Journal of New Music Research*, vol. 25, no. 1, pp. 1–30, 1996.

[20] E. Pampalk, P. Herrera, and M. Goto, "Computational models of similarity for drum samples," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 408–423, February 2008.

[21] A. Danielsen, C. H. Waadeland, H. G. Sundt, and M. A. Witek, "Effects of instructed timing and tempo on snare drum sound in drum kit performance," *Journal of the Acoustical Society of America*, vol. 138, pp. 2301–2316, October 2015.

[22] J. P. Bello *et al.*, "A tutorial on onset detection in music signals," *IEEE Transactions on speech and audio processing*, vol. 13, pp. 1035–1047, September 2005.

[23] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: an open-source library for sound and music analysis," in *21st ACM International Conference on Multimedia*, pp. 855–858, ACM, October 2013.

[24] D. Moffat, D. Ronan, J. D. Reiss, *et al.*, "An evaluation of audio feature extraction toolboxes," in *18th International Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway*, November 2015.

[25] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of mpeg-7," in *ICMC: International Computer Music Conference*, pp. 166–169, September 2000.

[26] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America*, vol. 130, pp. 2902–2916, November 2011.

[27] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, pp. 2579–2605, November 2008.

[28] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in neural information processing systems*, pp. 857–864, December 2003.

[29] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, pp. 2319–2323, December 2000.

[30] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, pp. 2323–2326, December 2000.

[31] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications.* Springer Science & Business Media, August 2005.

[32] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856, September 2002.

[33] A. Wilson and B. Fazenda, "Variation in multitrack mixes: analysis of low-level audio signal features," *Journal of the Audio Engineering Society*, vol. 64, pp. 466–473, August 2016.

[34] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: a structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, pp. 22–30, March 2011.

[35] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, October 2011.

[36] S. Bech and N. Zacharov, *Perceptual audio evaluation: theory, method and application.* Wiley Online Library, April 2006.

[37] "Recommendation itu-r bs.1534-1: Method for the subjective assessment of intermediate quality level of coding systems," 2003.

[38] N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, "Web Audio Evaluation Tool: A browser-based listening test environment," in *12th Sound and Music Computing Conference*, July 2015.

[39] A. Mehrabi, S. Dixon, and M. Sandler, "Vocal imitation of percussion sounds: On the perceptual similarity between imitations and imitated sounds," *Plos one*, vol. 14, p. e0219955, July 2019.

[40] J. M. Taylor, "Kendall's and spearman's correlation coefficients in the presence of a blocking variable," *Biometrics*, pp. 409–416, June 1987.

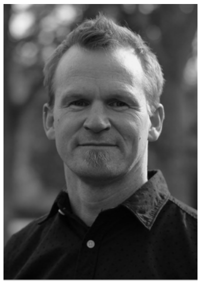
[41] M. Brueckl, "Statistische verfahren zur ermittlung der urteileruebereinstimmung," *Altersbedingte Veraenderungen der Stimme und Sprechweise von Frauen, Berlin: Logos*, vol. 88, p. 103, October 2011.

[42] M. Brueckl, F. Heuer, and M. M. Brueckl, "Package 'irrna'," April 2018.

[43] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," *IEEE Signal Processing Magazine*, vol. 26, May 2009.

THE AUTHORS

Jordie Shier



Kirk McNally

George Tzanetakis

Ky Grace Brooks

Jordie Shier is currently a pursuing a master's degree in Computer Science and Music at the University of Victoria, supervised by Professor George Tzanetakis and Professor Kirk McNally. His master's thesis research is focussed on the development of intelligent music production tools using techniques from the field of music information retrieval. His research is funded by a Canadian NSERC fellowship. He completed a B.Sc in Combined Computer Science and Music at the University of Victoria in 2017 and was a 2016/17 recipient of the Jamie Cassels Undergraduate Research Award for his work on the computational analysis of percussion sounds.

Kirk McNally is Assistant Professor of Music Technology in the School of Music at the University of Victoria, Canada. He is the program administrator for the undergraduate combined major program in music and computer science and the graduate program in music technology. Kirk is a sound engineer who specializes in popular and classical music recording, and new music performances using electronics. He has worked in studios in Toronto and Vancouver, with artists including REM and Bryan Adams and with ensembles including the Aventa Ensemble and the Victoria Symphony. His research and creative work has been supported by the Deutscher Akademischer Austausch Dienst (DAAD), the Canada Council for the Arts, the Banff Centre for Arts and Creativity and the Social Sciences Humanities Research Council of Canada (SSHRC).

George Tzanetakis is a Professor in the Department of Computer Science with cross-listed appointments in ECE and Music at the University of Victoria, Canada. He is the Canada Research Chair (Tier II) in the Computer Analysis of Audio and Music and received the Craigdarroch research award in artistic expression at the University of Victoria in 2012. In 2011 he was Visiting Faculty at Google Research. He received his PhD in Computer Science at Princeton University in 2002 and was a Post-Doctoral fellow at Carnegie Mellon University in 2002-2003. His research spans all stages of audio content analysis such as feature extraction, segmentation, classification with specific emphasis on music information retrieval. He is also the primary designer and developer of Marsyas an open source framework for audio processing with specific emphasis on music information retrieval applications. His pioneering work on musical genre classification received a IEEE signal processing society young author award and is frequently cited. More recently he has been exploring new interfaces for musical expression, music robotics, computational ethnomusicology, and computer-assisted music instrument tutoring. These interdisciplinary activities combine ideas from signal processing, perception, machine learning, sensors, actuators and human-computer interaction with the connecting theme of making computers better understand music to create more effective interactions with musicians and listeners.

Ky Grace Brooks is a PhD Candidate at McGill University, where they are studying the role of tacit knowledge in the co-construction of gender and professional performativities in audio engineering. They are also an experimental musician and live sound engineer.