GROOVE KERNELS AS RHYTHMIC-ACOUSTIC MOTIF DESCRIPTORS

Andy M. Sarroff
Dartmouth College
Department of Computer Science
sarroff@cs.dartmouth.edu

Michael Casey
Dartmouth College
Departments of Computer Science and Music
michael.a.casey@dartmouth.edu

ABSTRACT

The "groove" of a song correlates with enjoyment and bodily movement. Recent work has shown that humans often agree whether a song does or does not have groove and how much groove a song has. It is therefore useful to develop algorithms that characterize the quality of groove across songs. We evaluate three unsupervised tempo-invariant models for measuring pairwise musical groove similarity: A temporal model, a timbre-temporal model, and a pitch-timbre-temporal model. The temporal model uses a rhythm similarity metric proposed by Holzapfel and Stylianou, while the timbre-inclusive models are built on shift invariant probabilistic latent component analysis. We evaluate the models using a dataset of over 8000 real-world musical recordings spanning approximately 10 genres, several decades, multiple meters, a large range of tempos, and Western and non-Western localities. A blind perceptual study is conducted: given a random music query, humans rate the groove similarity of the top three retrievals chosen by each of the models, as well as three random retrievals.

1. INTRODUCTION

The propensity to move to music in a particular way is widespread and fundamental to our experience of music listening and enjoyment. Anyone who has spontaneously bopped their head, clapped their hands, jumped the pogo, swayed their cigarette lighter in the air, or tapped their fingers or toes to music has shared this common experience of near-involuntary musical response. Yet this aspect of music has been little studied in music information retrieval.

The phenomenon has been variously described as flow [3], sensorimotor synchronization [9], feel [16], and groove [8, 11, 12, 16]. Although related, the concept of groove is different from beat, which is the property of a predictable underlying periodic pulse [17]. The degree of groove is correlated with the degree to which the music induces the desire to move rather than the location and frequency of periodic entrainment.

We propose a new algorithm that extracts *groove kernels*—underlying audio patterns that correlate with the propensity to move. We use these features to measure groove similarity between pieces of music. We define groove similarity as that aspect of the sound pattern that induces, within a subject, the desire to move *in the same way*.

We conducted a groove similarity experiment using human subjects. The experiment evaluated three automatic groove extraction and similarity algorithms: an extant tempo-invariant rhythm similarity measure [6], and two versions of a proposed tempo-and-shift invariant groove kernel extraction system. The proposed system is inspired by the work of [18,20] which we extend with a full groove-oriented system architecture.

Our evaluation with human subjects used a diverse dataset of real-world audio files. Results show that our system retrieves music that corresponds more closely to human judgements about groove similarity than random baselines. In summary, the primary contributions of this work are:

- a new *groove kernel* feature based on shift and time-scale invariant Probabilistic Latent Component Analysis,
- a new dataset consisting of over 8,000 musical audio tracks in 10 contrasting genres,
- evaluation of three extraction and retrieval systems and two random baselines against human groove similarity judgments.

To the best of our knowledge, this is the first work to detail an unsupervised system architecture for groove features and experimentally evaluate it on human subjects. The following section provides background and motivation. Section 3 describes how groove kernels are extracted from audio files. Experimental evaluation is presented in Section 4, followed by concluding remarks in Section 5.

2. BACKGROUND

2.1 Groove

Underlying the operational character of groove is the psychological sensorimotor synchronization of auditory stimuli and human movement. In their study Janata et al. [9] showed that subjects strongly agreed that groove was described by the extent to which music induces movement, a positive affect due to such proclivity, and a feeling of being part of the music. From a sensorimotor perspective [12] defined groove as "wanting to move some part

of the body in relation to some aspect of the sound pattern". Other studies consider timing deviations [1], or temporal discrepancies [11], with respect to precise metronomic timing as the source of the groove, relating these to expressiveness and proclivity for motion. Pressing describes groove, or feel, as a "firmly structured temporal matrix" [16]. It is a temporal foundation and an emergent phenomenon formed out of concurrent recurring pulses (a stable sense of tempo), perception of a cycle of time that lasts for 2 or more pulses, and is effective in engaging synchronization of bodily movement.

Following [16] we take the position that groove induces characteristic responses in subjects and these responses stem from specific repeated acoustic patterns. Substantially different patterns induce different tendencies of motion, therefore the feel or the groove is different. Music that grooves is characterized by strong repetition. Therefore also following [16], we expect to observe a foundational "temporal matrix" that expresses the acoustic pattern corresponding to a particular groove at the time scale of roughly two bars. We hypothesize that such foundational patterns are invariant to shifts in time (i.e. within a song) and shifts of tempo (i.e. between songs).

2.2 Beat, Meter, Rhythm

To express invariance to shifts in tempo the description of groove must be normalized to the concept of beat. Alignment of a temporal matrix to the beat is not enough for comparisons between musical excerpts. There must also be a way to normalize for the phase of a temporal pattern with respect to beat hierarchy, or meter. There are two approaches to this problem: bar extraction and circular shifting of the temporal matrix. If we wish to represent groove as a multi-bar pattern, then we must rely on circular shifting.

Holzapfel et al. [6] addresses tempo invariant representations of rhythm at multiple time scales, therefore characterizing multi-scale rhythm. This work is unique in that it provides a scale-invariant song-level rhythm descriptor for music. Holzapfel et al. show that music with similar albeit complex rhythmic structure may be successfully categorized, even when the tempos are rather different. Therefore we see their algorithm as a candidate representation for groove similarity.

2.3 Rhythm Retrieval and Classification

While groove retrieval has not been explicitly treated by the music information retrieval community, rhythm classification and retrieval has been studied in recent years. Rhythm similarity metrics typically extract a rhythm descriptor that exhibits tempo-invariant properties. For instance [5] measures pairwise rhythm similarity using beat spectra based upon beat-synchronous low level features. Pattern segmenting is used in conjunction with dynamic time warping of acoustic features for pairwise rhythm similarity in [13].

The annotation of a large-scale rhythm-based dataset is expensive. Several authors have leveraged the music recordings available at the Ballroom Dancer's website [1] which have tempo and genre annotations. With this dataset authors have presented the results of genre classification tasks using rhythm descriptors such as amplitude envelopes of bar/beat synchronous features [4]; log-scale autocorrelation of onset strength signals [10]; fluctuation patterns [15]; and spectral rhythm patterns [14]. The success of many of these approaches is augmented when tempo metadata from the dataset is included. Hence the experimental results reported often reflect a semi-supervised approach.

2.4 Shift-Invariant Representation

To extract the most salient repeated aspects of the music we use shift-invariant probabilistic latent component analysis (SI-PLCA) [18]. A convolutional variant of non-negative matrix factorization (NMF), SI-PLCA places NMF in an explicitly Bayesian framework and extracts time-frequency components that are stable to shifts in time or frequency.

Our work focuses on time-shift invariant PLCA. Given a nonnegative matrix V, time-shift invariant PLCA factorizes V such that $V \approx \sum_k z_k W_k * \mathbf{h}_k$, k is an index to the factor components, Z is a diagonal matrix containing mixing coefficients, and $*$ is the convolutional operator. In our models we extract one component. Hence $z = 1$, W is a two-dimensional matrix, and $\mathbf{h}$ is a vector. We refer to W as a kernel and $\mathbf{h}$ as an activation function locating the kernel at multiple positions in a track.

Weiss and Bello [20] used SI-PLCA to evaluate song structure segmentation in a Beatles data set. They employed chroma features to extract multiple phrase-level blocks within songs whereas we use constant-Q spectral and cepstral features to extract single rhythmic kernels at the bar level using different sparseness constraints. Finally, we assessed kernels in a groove similarity task with a large and diverse dataset using human evaluators.

3. SYSTEM ARCHITECTURE

The groove kernel is built in four stages. In the first stage, bar and beat detection is performed. In the second stage, beat-synchronous features are extracted and bar/beat activation templates are generated. The third module estimates meter. The fourth stage extracts a shift-invariant groove kernel. The overall architecture is depicted in Figure 1. The details of each stage are described below.

3.1 Beat and Bar Tracking

Given a discrete time audio signal, we perform bar and beat tracking using the Queen Mary bar and beat tracker [2] reported in [2, 19]. The meter of the audio is a required input parameter for the bar tracker. Since we do not know the meter of a given musical audio file, we run the bar tracker twice: once assuming 3/4 meter and again assuming 4/4 meter. We note that changing the value of the input parameter to the beat/bar tracker does not affect the estimated

[1] http://www.ballroomdancers.com/

[2] Available as a Vamp plugin at http://isophonics.net/QMVampPlugins.

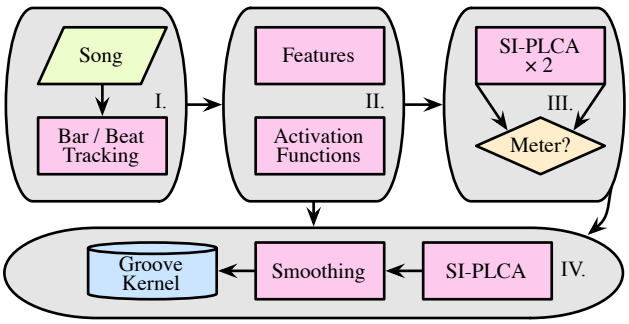

Figure 1. Overview of system architecture.

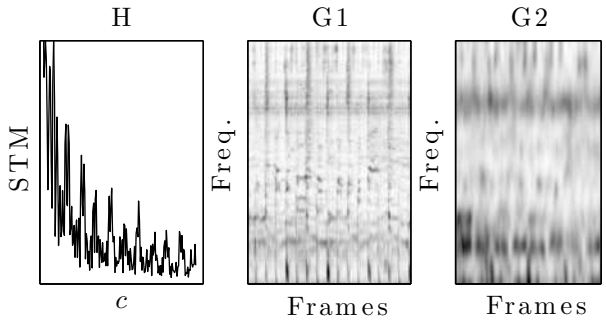

Figure 2. Left: *H* descriptor. The x axis plots the scale coefficient, with $1 \leq c \leq 100$, averaged across all frames. Middle: *G1* descriptor. Right: *G2* descriptor.

locations of the beats, only the indices that represent estimated bar onsets. The Queen Mary beat tracker has a reported accuracy of 73.6% when metrical level is not taken into account. The downbeat detector has a reported accuracy of 52.6%. Beat and downbeat tracking is an open problem and these results are comparable to the state of the art.

3.2 Beat Synchronous Features and Activation Templates

We extract frequency domain beat synchronous features. In this work, we use two feature types—the Constant Q Fourier Transform (CQFT) and the Low-Quefrency Constant Q Fourier Transform (LCQFT). The CQFT is computed by applying a log frequency-spaced filterbank to the Short Time Fourier Transform (STFT) of the audio signal. The LCQFT is computed by transforming the CQFT of the audio signal to the cepstral domain, applying a low-pass lifter, and inverting the signal back to the log frequency domain, analogous to MFCCs.

There are several parameters to choose when extracting the low level features. In this work, our audio has a sample rate of 22050 Hz; we use 2048-point FFTs over hamming-windowed frames of audio. The hop size is dynamically determined based upon estimated beat locations. The duration of each estimated beat is allocated 16 feature frames. The CQFT is computed using 24 bands per octave beginning at the approximate frequency of the musical note C2, yielding 178 CQFT coefficients per frame. We use 15 lower cepstral coefficients of the CQFT to compute the LCQFT (the first cepstral coefficient is ignored). A linear pre-emphasis is placed over the frequency channels to give higher weight to higher frequencies. This weighting helps SI-PLCA avoid placing too much probability on the lower frequencies.

The CQFT preserves pitch information while reducing spectral resolution at higher frequencies. The liftering stage of the LCQFT effectively removes much of the pitch information from the signal while retaining the timbre information. Hence the CQFT-based model is a pitch-timbre-temporal model, while the LCQFT-based model is a timbre-temporal model. We refer to the respective models as *G1* and *G2* in the rest of this paper.

We generate several activation templates based upon the estimated locations of beats and bars. The templates are used as priors over the activation function $\mathbf{h}$ for meter estimation and groove kernel extraction. The amplitudes of the spikes sum to one. A duple meter and a triple meter activation temple are generated without regard to the bar locations. These have a spike every second and third beat, respectively. The templates are used for meter estimation. A duple and triple meter activation template are also generated for groove kernel extraction. The duple-meter activation template has a spike every fourth beat and the triple meter activation template has a spike every third beat. These templates are organized such that the first spike is centered on the estimated location of the first bar line.

3.3 Meter Estimation

We do not know *a priori* what the meter of a given song is. To extract an effective groove representation in the following stage, we set the size of the kernel based upon a meter assumption. In this stage, we make a decision about the meter assumption using the log probabilities of two SI-PLCA models.

The meter estimation activation templates are given as prior probabilities to two independent SI-PLCA models: one with a triple, and the other with a duple meter assumption. The triple-meter model has a kernel window size of 96 frames (2 bars in 3/4). The duple-meter model has a kernel window size of 128 frames (2 bars in 4/4). The triple and duple models are run until convergence, with updates to the activation functions allowed. There are no sparsity constraints imposed on the model optimizations. However, since the initial activation functions are sparse, the final activation functions are also sparse. The meter of the song is chosen according to whichever model has the highest log probability after convergence.

3.4 Groove Kernel

Once the meter has been chosen, we extract a groove kernel using a final stage of SI-PLCA. We provide a new initial activation template $\mathbf{h}$ to the model in which there is an impulse every 4 or 3 beats, based upon the assumed meter. Note that a duple meter model now has activations every

4 beats, instead of 2. We also use the bar estimations produced in the second stage to center the activations at the onsets of 4/4 or 3/4 bars. The window size is set to two bars.

Weiss and Bello [20] have suggested that the optimal window size and meter may be learned by setting a sloping prior over an initial $\mathbf{h}$. We tried this approach using varying initializations of $\mathbf{h}$, slope degrees, initial window sizes, and sparsity parameters. In informal listening tests we found that our method provided better qualitative results when the bar and beat tracker was accurate.

Once a groove kernel has been extracted we smooth it along the time axis using a gaussian window. In this paper *G1* has no smoothing and *G2* is smoothed with a gaussian window having a standard deviation of 1 frame. Since our features have 16 frames per beat this window places approximately 95% of the window over 4 frames, or 1/16 note.

We do not know whether the phase of the groove kernel is aligned with respect to a latent two-bar groove structure of the music. Therefore for every groove kernel we enter a zero-phase and a circularly-shifted 1/2-phase version into our database.

4. EXPERIMENTS

4.1 Dataset

We built a dataset consisting of thousands of songs to evaluate the algorithms presented in this paper. All data is publicly available and we will provide the aggregate dataset and all associated metadata upon request. The data collection steps are summarized below.

We used the Echo Nest developer's API [3] to construct a list of 10,000 song titles across 10 genres and 10,000 unique artists. We began by querying the top styles in Echo Nest's database. An Echo Nest "style" is a search term associated with artists. Styles are essentially genres; the top ranked styes are those that Echo Nest believes yield the strongest search results. We will refer to Echo Nest styles as genres hereafter. We handpicked 10 genres from the highest ranked members of the list that we associated with having groove and variety. Table 1 shows the genres we selected, along with their Echo Nest rank.

For each genre we queried 1000 unique artists that were also cross-indexed with the 7digital [4] database, ranked by genre relevance. For each artist we queried 1 unique song that was in the 7digital database, ranked by Echo Nest's highest "danceability" estimation.

7digital is a commercial music distribution service that maintains .mp3 previews for most of the songs in their catalogue. We downloaded all previews in our list from the 7digital website. While sampling the dataset we discovered anomalous files. We filtered these out, resulting in 8249 unique song/artist clips each between 30 and 60 seconds long. We believe that the dataset dually exhibits a wide representation of groove and low redundancy.

[3] http://developer.EchoNest.com/

[4] http://us.7digital.com/

Rank and Genre										
1	rock	2	electronic	3	hip hop	6	jazz	14	pop	
17	reggae	19	funk	88	latin jazz	168	world	179	country	

Table 1. Echo Nest ranks and genres used in this work.

4.2 Models

We investigated three models, designated *H*, *G1*, and *G2*. *H* yields a temporal rhythm descriptor. *G1* yields a pitch-timbre-temporal groove kernel. *G2* yields a timbre-temporal groove kernel.

The *H* model is the scale invariant rhythm descriptor presented by Holzapfel and Stylianou in [7]. *H* computes multiple Direct Scale Transforms (DSTs) on the autocorrelation function of an Onset Strength Signal. We follow the procedure outlined in [7]. The DST is computed over a range of scale coefficients. The value of the maximum coefficient is denoted as C. Holzapfel and Stylianou show that the optimal value of C is related to the source material, but a value of $C > 80$ achieves nearly constant accuracy in their rhythm similarity tasks. The *H* model sets $C = 100$. The final descriptor is the average of the scale transform magnitudes across frames.

The other two models are *G1* and *G2*. Their architecture and parameterization are described in Section 3. Note that the key differences between *G1* and *G2* are that *G1* uses CQFT features. *G2* is built with LCQFT features and has smoothing over the groove kernel.

Figure 2 graphically depicts three extractions from the same song clip using *H* (left), *G1* (middle), and *G2* (right). Observe that the *H* descriptor is a vector of Scale Transform Magnitude (STM) against a range of scaling coefficients (denoted c). *G1* and *G2* exhibit different images even though they are extracted on the same audio clip. *G1* has finer-grained detail in the temporal domain and a sustained tone with harmonics in the upper third of the image. The rhythmic structure is apparent in *G2*, but the tonal and temporal detail has been smoothed.

4.3 Methods

A subset of 100 musical queries—10 from each genre—were randomly selected from the dataset. For each query and each model the top-3 nearest neighbors were selected (excluding the same song), as measured by cosine similarity. Retrievals for the *G2* model were additionally restricted to a have an estimated tempo difference of 8 BPM to limit the range of tempo variation.

There were two sets of random retrievals. The *R1* retrieval set has 3 songs chosen at random for each query. Each retrieval in the *R2* set has an estimated tempo difference from its associated query of less than or equal to 8 BPM. Tempos were estimated by computing the median beat onset differences derived from the bar and beat tracker. Random retrieval sets were not restricted by genre.

There were two types of participants: solicited and anony-

mous. Solicited participants were paid if they completed the entire experiment. Both types of participants were presented the same web-based experiment interface. Participants were randomly assigned one of ten genre-based test subsets. A test subset consists of 10 same-genre queries and 5 retrieval sets. We collected 2,436 ratings from solicited participants and 236 ratings from anonymous participants. There were 56 unique human evaluators that participated in our experiment.

The experiment required that participants utilize a quiet listening room or headphones. Participants were presented with the following definition of groove [9]: "The groove is that aspect of the music that induces a pleasant sense of wanting to move along with the music."

We are unaware of any studies in the literature on the perception of groove *similarity*. We therefore asked participants to consider the similarity of groove based upon the given definition. The experimental interface further stated, "Please try to avoid judging groove similarity based upon song genre. For instance, you may find that two songs are from different genres, but you would move your body in a similar way to them. You should rate these songs as having high groove similarity."

Each participant was presented query-retrieval pairs from their test subset in random order. The audio clips were approximately 5 seconds in duration, corresponding to the expected duration of a two bar motif. They were asked to rate groove similarity on a coarse scale using radio buttons having the labels "Not Similar", "Somewhat Similar", and "Very Similar". These ratings were later assigned numerical values from the set $\{0, 1, 2\}$. We denote these as "coarse" ratings. Participants were also asked to rate the groove similarity of each pair on a fine scale with a slider. The slider had a range of $[0, 100]$, but the slider's numerical value was not exposed to the user. We denote the ratings as "Fine".

A participant was required to listen to each pair of audio clips at least once and assign ratings before moving to the next comparison. Multiple listens were permitted. We note that the experimental design was modeled after the MIREX Audio Music Similarity and Retrieval [5] evaluation procedure. One minor difference is that the Mirex Audio Similarity task asks its evaluators to rate the top-5 ranked songs per query and model. Due to limited human resources, we restricted the retrieval space to top-3.

4.4 Results

Figure 3 shows the mean coarse and fine ratings per retrieval set. The red cross hairs show standard error. Table 2 shows the results of pairwise t-tests between appropriate baselines and models. Note that *G2* may only be directly compared with *R2*; these were the retrieval instances where the search space was restricted by tempo difference.

The first thing that we notice is that participants were pessimistic about the groove similarity of query-retrieval pairings. The mean coarse value across all ratings was 0.780; the mean fine rating was 40.909. Users were more likely to rate a pair of songs as being not similar or somewhat similar than very similar. The songs in the dataset spanned a range of 10 base genres. Several participants expressed that they had difficulty cognitively separating genre and preference from groove. Indeed, Janata et al. have shown that enjoyment is correlated with groove [9]. We are not aware of a study that evaluates correlation between genre and groove similarity.

[5] http://www.music-ir.org/mirex/wiki/Audio_Music_Similarity_and_Retrieval

	R1-H	*R1-G1*	*R2-G2*
Coarse	$\mathbf{4.52 \times 10^{-5}}$	**0.0038**	0.1160
Fine	$\mathbf{1.66 \times 10^{-4}}$	**0.0094**	**0.0142**

Table 2. Pairwise t-test p-values. Boldface indicates a p-value less than 0.05.

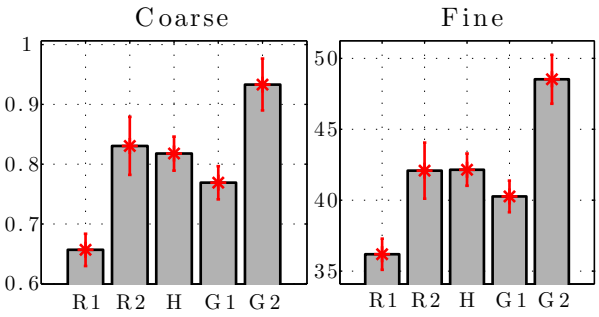

Figure 3. Mean coarse and fine ratings by algorithm. The red cross hairs show standard error about the mean.

Secondly we observe that all models retrieve groove-similar songs better than random selection when the retrieval space is unrestricted by tempo. We have learned from Janata et al. that humans are able to reliably detect the presence of groove. Our results support the hypothesis that humans may also reliably detect groove similarity.

We find that *G1* and *H* perform competitively. When adjusting for multiple comparisons using the Tukey-Kramer method each performs significantly better than *R1* (at 95% confidence), but they share similar statistical distributions with each other for coarse and fine ratings.

We notice that groove similarity ratings jump upward for random retrieval when the space is limited to an 8 BPM tempo difference from the query. Humans are more likely to rate two arbitrary songs to have similar groove if they are close in tempo.

The only model that was evaluated with a restricted tempo space was *G2*. As can be seen in Figure 3 and Table 2, this model performed significantly better than the tempo-restricted random set on fine evaluations. We do not know whether the increased performance of *G2* is due to a (pitch-free) low-level feature or the gaussian smoothing of the groove kernel. Our intuition leads us to believe that smoothing had a significant impact. The kernels are fairly high-dimensional. By smoothing them, neighbors that were once distant due to fine differences in temporal structure become less distant (cf. Figure 2).

5. CONCLUSIONS

Groove is associated with the often pleasurable induction of bodily movement to music. There are an increasing number of rhythm similarity and classification algorithms in the literature, yet groove encompasses a higher-level construct involving sensorimotor interaction stemming from repeated acoustic patterns. We presented a new groove kernel feature based on shift and tempo invariance. We asked humans to evaluate the groove kernel and another rhythm similarity model in a groove similarity retrieval task using a diverse collection of real-world music recordings spanning 10 base genres.

The *H* and *G1* models give groove similarity rankings that are significantly better than random retrieval. The *G2* model performs significantly better than random retrieval when the retrieval space is limited to an 8 BPM absolute difference from the query.

We note that all three models—the temporal *H* model and our proposed timbre-inclusive SI-PLCA based models—are constructed in an unsupervised manner. Building human-annotated collections of music is expensive. Hence there is higher value associated with models that do not rely on human annotation.

Our models rely on beat-synchronous features derived from the automatically estimated bar and beat estimations. As noted in Section 3 beat and downbeat estimation is not a solved problem. We may assume that there is error in the estimated beat and bar locations. Unfortunately, this error is necessarily propagated forward through every stage of the groove kernel models. We expect that our proposed models will perform better as beat and downbeat detection improves.

The groove kernel activation templates were restricted to 3/4 and 4/4 meters. While the dataset included a large selection of world music, it is possible that the learned kernels did not fit a significant portion of the dataset. We expect that the algorithm could be improved with enhanced meter detection.

The experimental design did not allow for a direct comparison between the *H* and *G2* or *G1* and *G2* methods. We therefore cannot draw conclusions regarding the impact of the tempo-restricted space on these methods. We also cannot state conclusively the contribution that smoothing has on the *G2* model since this effect was not studied.

Our human subject study had a limited number of human participants with respect to the number of song queries. There were 100 queries each associated with 3 models and two random baselines. An improved study would include the effects of pairwise perceived genre similarity, song preference, and other potential biases. Further investigation is needed into the relationship between *how much* groove is perceived and groove similarity. Future work will include a larger scale human evaluation with the intent to address these important issues.

To the best of our knowledge, this paper provides the first human-based groove similarity retrieval task. Experimental results suggest that the groove kernel presents a promising direction for exploration of groove metrics.

6. ACKNOWLEDGMENTS

This project has been supported by a Google Faculty Research Award and by a Neukom Institute for Computational Science Graduate Fellowship.

7. REFERENCES

[1] J.A. Bilmes. Timing is of the essence: perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, Massachusetts Institute of Technology, 1993.

[2] M.E.P. Davies and M.D. Plumbley. A spectral difference approach to downbeat extraction in musical audio. In *Proc. EUSIPCO*, 2006.

[3] O. de Manzano, T. Theorell, L. Harmat, and F. Ullen. Psychophysiology of flow during piano playing. *Emotion*, 10:301–311, 2010.

[4] S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *Proc. ISMIR*, volume 5, 2004.

[5] J. Foote, M. Cooper, and U. Nam. Audio retrieval by rhythmic similarity. In *Proc ISMIR*, volume 3, pages 265–266, 2002.

[6] A. Holzapfel and Y. Stylianou. A scale transform based method for rhythmic similarity of music. In *Proc. ICASSP*, pages 317–320. IEEE, 2009.

[7] A. Holzapfel and Y. Stylianou. Scale transform in rhythmic similarity of music. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):176–185, 2011.

[8] V. Iyer. Embodied mind, situated cognition, and expressive microtiming in african-american music. *Music Perception*, 19:387– 414, 2002.

[9] P. Janata, S.T. Tomic, and J.M. Haberman. Sensorimotor coupling in music and the psychology of the groove. *Journal of Experimental Psychology: General*, 141(1):54–75, 2012.

[10] J.H. Jensen, M.G. Christensen, and S.H. Jensen. A tempo-insensitive representation of rhythmic patterns. In *Proc. EUSIPCO*, 2009.

[11] C. Keil and S. Feld. *Music grooves*. University of Chicago Press, Chicago, IL, 1994.

[12] G. Madison. Experiencing groove induced by music: Consistency and phenomenology. *Music Perception*, 24:201–208, 2006.

[13] J. Paulus and A. Klapuri. Measuring the similarity of rhythmic patterns. In *Proc. ISMIR*, volume 2, 2002.

[14] G. Peeters. Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1242–1252, 2011.

[15] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On rhythm and general music similarity. In *Proc. ISMIR*, volume 9, 2009.

[16] J. Pressing. Black atlantic rhythm: Its computational and transcultural foundations. *Music Perception*, 19:285–310, 2002.

[17] E. Scheirer. Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, pages 588–601, 1998.

[18] P. Smaragdis, B. Raj, and M. Shashanka. Sparse and shift-invariant feature extraction from non-negative data. In *Proc. ICASSP*, pages 2069–2072, 2008.

[19] A.M Stark, M.E.P Davies, and M.D. Plumbley. Real-time beat-synchronous analysis of musical audio. In *Proc. DAFx*, 2009.

[20] R.J. Weiss and J.P. Bello. Unsupervised discovery of temporal structure in music. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1240–1251, oct. 2011.