MODELING MUSICAL RHYTHM AT SCALE WITH THE MUSIC GENOME PROJECT

Matthew Prockup[+*], *Andreas F. Ehmann*[*], *Fabien Gouyon*[*], *Erik M. Schmidt*[*], *Youngmoo E. Kim*[+]
{mprockup, ykim}@drexel.edu, {fgouyon, aehmann, eschmidt}@pandora.com

[+]Drexel University, ECE Dept.
3141 Chestnut St.
Philadelphia, PA 19104

[*]Pandora Media, Inc.
2101 Webster Street
Oakland, CA 94612

ABSTRACT

Musical meter and attributes of the rhythmic feel such as swing, syncopation, and danceability are crucial when defining musical style. However, they have attracted relatively little attention from the Music Information Retrieval (MIR) community and, when addressed, have proven difficult to model from music audio signals. In this paper, we propose a number of audio features for modeling meter and rhythmic feel. These features are first evaluated and compared to timbral features in the common task of ballroom genre classification. These features are then used to learn individual models for a total of nine rhythmic attributes covering meter and feel using an industrial-sized corpus of over one million examples labeled by experts from Pandora® Internet Radio's *Music Genome Project*®. Linear models are shown to be powerful, representing these attributes with high accuracy at scale.

Index Terms— audio, signal processing, music information retrieval, rhythm, feature engineering, large-scale machine learning

1. INTRODUCTION

Rhythm is one of the fundamental building blocks of music, and perhaps the simplest aspect for humans to identify with. But constructing compact, data-driven models of rhythm presents considerable complexity even when operating on symbolic data (i.e., musical scores). This complexity is compounded when developing algorithms to model rhythm in acoustic signals for organizing a large-scale library of recorded music. Previous work has studied the general recognition of rhythmic styles in music audio signals, but few efforts have focused on the deconstruction and quantification of the foundational components of global rhythmic structures. This work focuses on modeling rhythm-related attributes of meter and "feel" (e.g., "swing") in music by designing targeted acoustic features that can accurately represent these attributes on more than one million expertly-labeled audio examples from Pandora® Internet Radio's *Music Genome Project*® (MGP)[1].

The fundamental components of rhythm are metrical structure, tempo, and timing [1]. There is a large body of prior work that attempts to estimate these components [2, 3, 4, 5], but in extracting only beats, tempo, and meter much of the rhythmic subtlety and feel is discarded. A mid-level representation known as the *accent signal* [6], which measures the general presence of musical events, is better suited to represent this rhythmic subtlety. However, the tempo, beat, and meter estimates are still beneficial, as they can provide important temporal context to rhythmic patterns derived from the accent signal. For example, the frequencies of periodicity in an accent signal can be used to infer beats per minute, and when normalized by an estimate of tempo, directly relate to musical note durations [7]. The accent signal can also be quantized and viewed in the context of beats or measures in order to capture discrete instances of rhythm patterns [8, 9]. In other work, Holzapfel introduced the Mellin Scale Transform as both a tempo-invariant and tempo-independent method for describing rhythmic similarity. Unlike previous methods, the transform achieves tempo-invariance by design rather then normalizing by a tempo estimate [10].

Most of the previous work in capturing rhythm has relied on evaluation through the classification of a generalized musical style or genre, while simultaneously focusing on specific aspects of rhythm in the feature design. Evaluation is usually performed with the ballroom dance style dataset [11], which more precisely represents rhythm than a dataset that is labeled with basic genre. However, this remains a high-level approach with little regard to the meaning of the specific aspects of rhythm inherent in the music. As a result, researchers have started to overfit and exploit phenomena of the dataset rather than capture the attributes that relate more generally to music [12, 11]. Furthermore, work by Flexer demonstrates that general music similarity requires the context of many different factors outside of just rhythm [13]. While it is possible to argue that certain features may be capturing components of rhythm, the contextual complexities in the style labels make it difficult to infer meaning. This motivates the need for a more strict and concrete evaluation of rhythm features and their contributions to specific rhythmic components.

2. APPROACH

In this work, we seek to capture rhythmic attributes automatically in music audio signals. We designed and implemented a set of deterministic rhythmic descriptors that are tempo-invariant and represent specific elements of the rhythmic attributes. The descriptors are first compared and benchmarked to previous work with the widely-used ballroom style classification task. A large-scale evaluation is then performed using a set of linear machine-learning models to learn the presence of the meter and rhythmic feel components individually. This evaluation will show the descriptors' effectiveness at capturing each rhythmic attribute in music audio signals at scale.

The targeted attributes are compositional constructs, such as the meter, or well-defined components of the musical feel, such as the presence of swing. Namely we focus on the following 9 rhythmic attributes:

- **Cut-Time Meter** contains 4 beats per measure with emphasis on the 1^{st} and 3^{rd} beat. The tempo feels half as fast.

[1]"Pandora" and "Music Genome Project" are registered trademarks of Pandora Media, Inc. `http://www.pandora.com/about/mgp`

- **Triple Meter** contains groupings of 3 with consistent emphasis on the first note of each grouping. ($\frac{3}{4}$, $\frac{3}{2}$, $\frac{3}{8}$, $\frac{9}{8}$)
- **Compound-Duple Meter** contains 2 or 4 sub-groupings of 3 with emphasis on the 2^{nd} and 4^{th} grouping. ($\frac{6}{8}$, $\frac{12}{8}$)
- **Odd Meter** identifies songs which contain odd groupings or non-constant sub-groupings. ($\frac{5}{8}$, $\frac{7}{8}$, $\frac{5}{4}$, $\frac{7}{4}$, $\frac{6}{4}$, $\frac{9}{4}$)
- **Swing** denotes a longer-than-written duration on the beat followed by a shorter duration. The effect is usually perceived on the 2^{nd} and 4^{th} beats of a measure. (1 . . 2 . a 3 . . 4 . a)
- **Shuffle** is similar to swing, but the warping is felt on all beats equally. (1 . a 2 . a 3 . a 4 . a)
- **Syncopation** is confusion created by early anticipation of the beat or obscuring meter with emphasis against strong beats.
- **Back-Beat Strength** is the amount of emphasis placed on the 2^{nd} and 4^{th} beat or grouping in a measure or set of measures.
- **Danceability** is the utility of a song for dancing. This relates to consistent rhythmic groupings with emphasis on the beats.

Previous work has looked at identifying musical meter. However, emphasis was placed on distinguishing duple versus triple in a more general sense rather than identifying the true meter, which has an important function in the context of rhythmic style. Because focus is placed on meter differentiation, we target *cut-time*, *triple*, *compound-duple*, and *odd* meters and ignore widely shared ones such as *simple-duple* ($\frac{2}{4}$, $\frac{4}{4}$). Rhythmic feel has also been studied, but mostly in the context of similarity. Individual components of the rhythmic feel are important in defining style. They are easily recognizable to a listener, but are sometimes difficult to quantify. In this work we seek to define and capture the the qualities of *swing*, *shuffle*, *syncopation*, *back-beat strength*, and *danceability*. The rhythmic component labels were defined and collected by musical experts on a corpus of over one million audio examples from Pandora® Internet Radio's *Music Genome Project*®(MGP). The labels were collected over a period of nearly 15 years and great care was placed in defining them and analyzing each song with a consistent set of criteria.

3. DESIGNING FEATURES FOR RHYTHM

In order to capture aspects of each rhythm label, a set of rhythm-specific features was implemented. The features are based on an *accent signal*, which measures the change of a music audio signal over time. High points of change denote the presence of a new musical event. The accent signal used is a variant of the SuperFlux algorithm [6] and is the half-wave rectified ($H(X) = \frac{X+|X|}{2}$) sum of frequency bands of a frequency smoothed (Eq. 1) constant-Q transform X_{cqt} of an audio signal (Eq. 2).

$$X_{cqt}^{max}[n,m] = \max(X_{cqt}[n, m-1:m+1]) \quad (1)$$

$$SF[n] = \sum_{m=1}^{m=M} H(X_{cqt}[n,m] - X_{cqt}^{max}[n-\mu, m]) \quad (2)$$

From the accent signal, an estimate of tempo is found. This was achieved through a hybrid of the standard inter-onset-interval (IOI) and autocorrelation function (ACF) methods that are widely used. The IOI method employs SuperFlux onset detection to create a histogram of inter-onset-distances. The ACF method is the autocorrelation of the accent signal. *Periodicity salience* is then found by summing across k harmonics and sub-harmonics of the ACF lag or the IOI Histogram distance (Eq. 3).

$$S_{\text{ACF}}[l] = \sum_{k=1}^{K} \text{ACF}[kl] + \sum_{k=2}^{K} \text{ACF}[\frac{1}{k}l] \quad (3)$$

Periodicity salience is then converted to a *tempogram* by transforming the onset distance or lag l in time to a tempo $\tau = \frac{60}{l}$bpm. A fusion tempogram $F_{TG}(\tau)$can be found by multiplying the individual tempograms (Eq. 4) [7].

$$F_{TG}(\tau) = S_{\text{ACF}}(\tau) \odot S_{\text{IOI}}(\tau) \quad (4)$$

A tempo estimate can then be found by taking the tempo related to the maximum peak in the fusion tempogram.

Using this accent signal and the tempo estimate, beat tracking is performed using dynamic programming [2]. This method was chosen for its ease of implementation, scalability, deterministic nature, and consistency of beat position estimation. Each rhythm feature described in the following sections are derived using a combination of the accent signal, tempograms, and beat estimates. In order to visualize each feature, a set of consistent style examples of Samba, Tango, and Jive from the ballroom dataset will be used. A canonical representation of these rhythms for drum set obtained from Tommy Igoe's *Groove Essentials* is shown in Figure 1 [14].

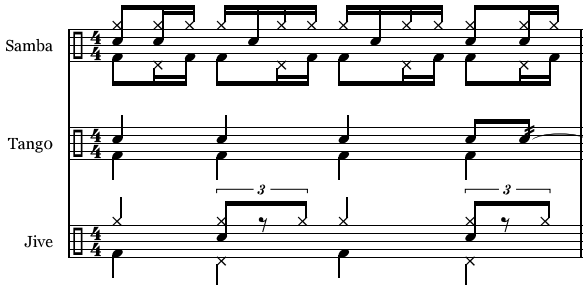

Figure 1: These patters define the Samba, Tango, and Jive rhythmic styles for drum set.

3.1. Beat Profile

The *beat profile* is a compact snapshot of the accent signal that takes advantage of the beat estimates. This is similar to the feature by Dixon [8], but it is simpler, deterministic, and free of human intervention. The accent signal between consecutive beats is quantized in time to 36 beat subdivisions. The beat profile features are statistics of each of those 36 bins over all beats. The beat profile distribution feature (BPDIST) is comprised of the mean of each beat profile bin (BPMEAN) and constrained such that that the collection of bins must sum to one. A set of beat profile features is shown in Figure 2.

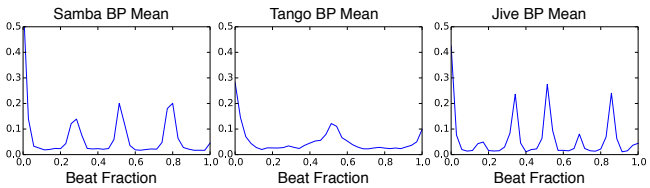

Figure 2: Examples of the BPMEAN Feature.

3.2. Tempogram Ratio

The *tempogram ratio* feature (TGR) uses the tempo estimate, similar to work by Peeters [7], to remove the tempo dependence in the tempogram. By normalizing the tempo axis of the tempogram by the tempo estimate, a fractional relationship to the tempo is gained. A compact, tempo-invariant feature is created by capturing the weights of the tempogram at musically related ratios relative to the tempo estimate. Examples of the tempogram and tempogram ratio features are shown in Figure 3.

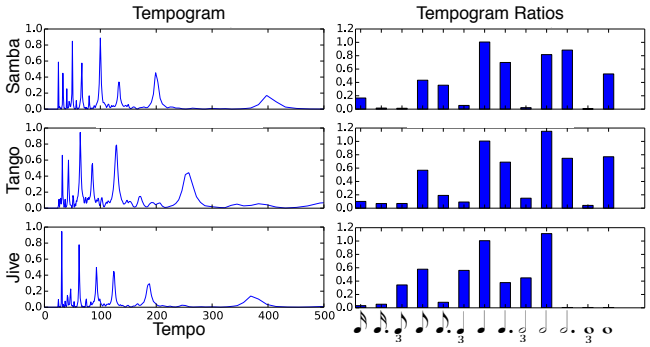

Figure 3: Examples of the TGR feature.

3.3. The Mellin Scale Transform

The *Mellin scale transform* is a scale invariant transform of a time domain signal. Similar musical patterns at different tempos are scaled relative to the tempo. The Mellin scale transform is invariant to that tempo scaling. It was first introduced in the context of rhythmic similarity by Holzapfel [10], around which our implementation is based. Scale-invariance comes at the cost of signal shift-invariance, so the normalized autocorrelation (Eq. 5) is used. The formulation for the Mellin scale transform $R(c)$ of discrete signals as a function of scale parameter c with autocorrelation lag time interval T_s is shown in Equation 6. The transform $R(c)$ is calculated discretely relative to the lag time interval T_s and window length time T_{up} (Eq. 7).

$$r'(l) = \frac{r(l) - \min\{r\}}{\max\{r\} - \min\{r\}} \text{ where, } r(l) = \sum_n x[n]\bar{x}[n-l] \quad (5)$$

$$R(c) = \frac{\sum_{k=1}^{\infty}[r'(kT_s - T_s) - r'(kT_s)](kT_s)^{1/2-jc}}{(1/2 - jc)\sqrt{2\pi}} \quad (6)$$

$$\Delta c = \frac{\pi}{\ln \frac{T_{up}+T_s}{T_s}} \quad (7)$$

The transform is calculated on autocorrelations of 8s widows with a 4s overlap. The song is summarized by the mean over time. An example of the scale transform feature (MELLIN) is shown in Figure 4. In order to exploit the natural periodicity in the transform, the discrete cosine transform (DCT) is computed. Median removal (by subtracting the local median) and half-wave rectifying the DCT creates a new feature that emphasizes periodicities. This new feature (MELLIN_D) is then normalized to sum to one. More about the Mellin scale transform can be found in [10, 15, 16].

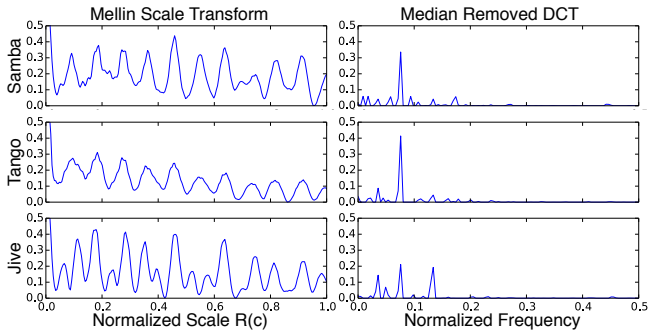

Figure 4: Examples of the MELLIN and MELLIN_D Feature.

3.4. Multi-band Representations

Each of the rhythm features described in sections 3.1 and 3.2 rely on a global estimates of beats, tempo and an accent signal. These features can be extended to multiple-band versions by using accent signals that are constrained to be within a set of specific sub-bands of the CQT: $(A_0, A_3]$, $(A_3, A_6]$, $(A_6, A_9]$. Using separate accent signals, the rhythmic features can relate to the different compositional functions of instruments in different frequency ranges.

3.5. Rhythmic Feature Evaluation

In order to evaluate and compare the new features, a set of general music information retrieval (MIR) classification tasks was performed on the ballroom dataset (8 ballroom dance styles, 698 instances, 523 instances with duple meter and 175 instances with triple meter). The rhythm features were used individually and in various aggregations with each feature dimension normalized from 0 to 1. Block-based Mel-Frequency Cepstral Coefficients (MFCC) are also used for comparison. Means and covariances of MFCCs are calculated across overlapping 6-second blocks. These block-covariances are further summarized over the piece by calculating their means and variances [17]. A simple logistic regression classifier was fit for 10 trials with a randomly shuffled 70:30 train:test split for each trial. A subset of these results is shown in Table 1.

Tempo-Invariant Feature	Dim.	Duple vs. Triple	Ballroom Style
BPDIST	36	0.849 ± 0.031	0.776 ± 0.035
BPDIST_M (multiband)	108	0.873 ± 0.016	0.794 ± 0.019
TGR	13	0.883 ± 0.024	0.747 ± 0.030
TGR_M (multiband)	39	0.952 ± 0.007	0.817 ± 0.022
MELLIN	230	0.956 ± 0.011	0.868 ± 0.010
MELLIN_D	230	0.936 ± 0.014	0.829 ± 0.018
MELLIN BPDIST_M TGR_M	377	$\mathbf{0.974 \pm 0.010}$	$\mathbf{0.917 \pm 0.018}$
MELLIN_D BPDIST_M TGR_M	377	0.959 ± 0.015	0.884 ± 0.019
MFCC	460	0.877 ± 0.016	0.511 ± 0.027
MFCC MELLIN BPDIST_M TGR_M	837	0.942 ± 0.018	0.743 ± 0.020
MFCC MELLIN_D BPDIST_M TGR_M	837	0.925 ± 0.017	0.707 ± 0.035
TG (tempo-variant)	500	0.962 ± 0.010	0.843 ± 0.011

Table 1: Ballroom dance style classification tasks results.

The tempogram (TG) feature shows state of the art performance on the ballroom dataset, which is evidence for the well-known class tempo-dependence [11]. Other features that are tempo-invariant perform similarly without exploiting the known class tempo-dependence of this dataset. Evidence of tempo-invariance in classification is shown by the confusion matrices for both the tempo-invariant and tempo-variant features. The tempogram (TG) confuses Jive (160-180bpm) with Waltz (78-98bpm), even though

they are very different stylistically. However, it cannot easily differentiate the exact 2:1 tempo ratio because both styles have energy at similar tempo multiples. Rumba (90-110bpm) and Jive show a similar error relationship. Conversely, MELLIN confuses Samba (96-104bpm) with Tango (120-140bpm) and ChaChaCha (116-128bpm), which do not overlap with Samba's tempo range. However, these three styles contain similarity in their rhythmic self-repetition, which is something the MELLIN feature is designed to capture. Furthermore, this lack of overlap makes Samba much easier to distinguish for the tempogram feature. This suggests that the rhythm features are representing something about the rhythmic characteristics, and not relying on tempo for discrimination.

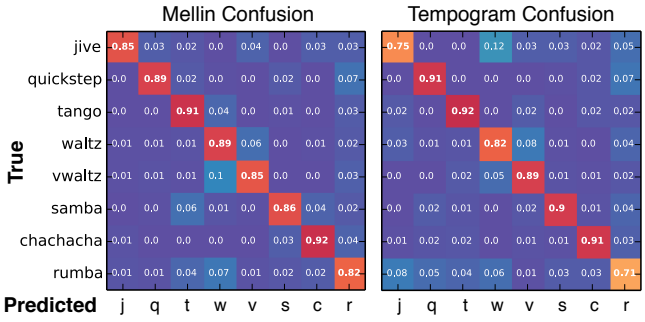

Figure 5: Confusion matrices of the Mellin Transform and Tempogram features for the ballroom dataset

4. PREDICTING RHYTHMIC ATTRIBUTES

In order to predict the rhythmic attributes from Section 2, stochastic gradient decent (SGD) was formulated for classification of the binary labels (log loss, logistic regression) and regression of continuous labels (least-squares loss, linear regression). The learning rate was tuned adaptively. The training data was separated on a randomly shuffled 70:30 train:test split with no shared artists between training and testing. Due to the size of the dataset, a single trial for each attribute is both tractable and sufficient. More on SGD can be found in [18]. Cut-time, triple, compound-duple, and odd meters along with the presence of swing, shuffle, and heavy syncopation are all binary attributes and are therefore formulated as classification tasks. Danceability and back-beat strength are continuous ratings and are formulated as regression tasks.

4.1. Results

The classification and regression results for each of the rhythm attributes are shown in Table 2. The binary classification tasks are evaluated using the area (AUC) under the receiver operating characteristic (ROC) curve. The regression results are evaluated with the R^2 metric.

The results show that the rhythm-motivated features are best able to capture the rhythm attributes when compared to the the timbre-motivated features. When both are used in combination, little improvement is gained. Timbre features alone can differentiate certain rhythmic attributes fairly well in some cases. For example, the cut-time meter is very common in the "country" genre and MFCC's are possibly picking up on the genre's similarly specific instrumentation rather than the rhythmic components. In all cases, the rhythm features are better than timbre alone, offering further

	AUC		Comp.					**R^2**	Back-
Features	Cut	Triple	Duple	Odd	Swing	Shuf.	Sync.	Dance	Beat
BPDIST	0.792	0.753	0.733	0.698	0.845	0.875	0.724	0.317	0.136
BPDIST_M (*B*)	0.864	0.807	0.772	0.756	0.871	0.886	0.745	0.412	0.301
TGR	0.645	0.759	0.804	0.728	0.795	0.840	0.658	0.317	0.136
TGR_M (*T*)	0.801	0.808	0.859	0.754	0.811	0.842	0.666	0.350	0.199
MELLIN (*S*)	0.810	0.916	0.945	0.840	0.868	0.914	0.743	0.452	0.269
MELLIN_D (*D*)	0.862	0.910	0.933	0.848	0.876	0.915	0.761	0.513	**0.425**
(S) (B) (T)	0.890	0.926	**0.949**	0.849	0.897	**0.921**	0.769	0.506	0.396
(D) (B) (T)	**0.899**	**0.924**	0.946	**0.862**	**0.902**	0.920	**0.770**	**0.515**	0.393
MFCC (*M*)	0.802	0.795	0.667	0.741	0.784	0.723	0.707	0.450	0.38
(M) (S) (B) (T)	0.899	**0.920**	**0.942**	0.843	0.897	**0.922**	**0.780**	**0.537**	0.464
(M) (D) (B) (T)	**0.904**	0.920	**0.942**	**0.861**	**0.903**	0.920	0.779	0.532	**0.468**

Table 2: The results for rhythm construct learning are shown. Both the AUC and R^2 metrics have a maximum value of 1.0 and lower bounds of 0.5 when predicting a random class and 0.0 when predicting the mean of the test labels.

proof that the rhythm features are learning something about the attributes they are targeting rather than their generalized correlation to a musical style.

Furthermore, it is seen among rhythm features that each have selected strengths. They tend to represent beat level versus measure level information and single-band (global) versus multiple-band (range-specific) information. When considering beat level versus measure level patterns, swing, shuffle, and syncopation are better represented by the beat profile features than the tempogram ratio features. This is because these rhythm attributes are defined on a local beat level, and the patterns within the beats have a specifically associated feel. Compound-duple and odd meter are better defined by tempogram ratios, which suggests that they have patterns that cannot be captured within a single beat. It is also seen that the Mellin representations are effective across beat-level and measure-level attributes, suggesting that they are able to capture both.

When looking at single-band versus multi-band features, the rhythm components and associated features that capture interplay between multiple instrument ranges are highlighted. Meter, syncopation, danceability and back-beats all rely on the emphasis of specific points in a measure. In the context of a performance, the use of multiple instruments may be used to highlight these differences in emphasis, which is captured in multi-band representations. Attributes that rely on global feel, such as a swing or shuffle, are not aided by the multi-band representations.

5. CONCLUSION

In this work, we outlined a set of tempo-invariant rhythmic descriptors that were able to distinguish rhythmic styles with state-of-the-art performance. We showed that they do not rely on exploiting the tempo-dependence of the ballroom dataset, which suggests that they are learning rhythmic characteristics and not simply tempo. A set of large-scale experiments was then performed to quantify and label a set of rhythmic meter and feel attributes using Pandora® Internet Radio's *Music Genome Project*®. From a musicology perspective, these rhythmic attributes are important in the makeup of a musical style. From this work, we gain insight into the meanings of rhythmic features as they relate to meter and feel when applying them to style recognition tasks in the future. In other future work, we plan to use more complex, scalable models such as Random Forests, Gradient Boosted Trees and stacked tree ensembles [19]. Similar to neural-network models, tree ensembles benefit from the ability to learn complex, non-linear mappings of the data.

6. REFERENCES

[1] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Computer music journal*, 2005.

[2] D. P. W. Ellis, "Beat Tracking by Dynamic Programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, Mar. 2007.

[3] J. Oliveira, F. Gouyon, L. Martins, and L. Reis, "IBT: A real-time tempo and beat tracking system," 2010.

[4] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *Audio, Speech, and Language Processing, IEEE Trans.*, vol. 14, no. 5, pp. 1832–1844, Sep. 2006.

[5] A. Klapuri and M. Davy, *Signal processing methods for music transcription*, 2006.

[6] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," *DAFx*, 2013.

[7] G. Peeters, "Rhythm Classification Using Spectral Rhythm Patterns." *ISMIR*, 2005.

[8] S. Dixon, F. Gouyon, and G. Widmer, "Towards Characterisation of Music Via Rhythmic Patterns," *ISMIR*, 2004.

[9] M. Prockup, J. Scott, and Y. E. Kim, "Representing Musical Patterns via the Rhythmic Style Histogram Feature," in *Proceedings of the ACM International Conference on Multimedia - MM '14*. New York, New York, USA: ACM Press, Nov. 2014, pp. 1057–1060.

[10] A. Holzapfel and Y. Stylianou, "Scale Transform in Rhythmic Similarity of Music," *Audio, Speech, and Language Processing, IEEE Trans.*, vol. 19, no. 1, pp. 176–185, Jan. 2011.

[11] F. Gouyon, "Dance music classification: A tempo-based approach," *ISMIR*, 2004.

[12] B. L. Sturm, "The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval," *Journal of New Music Research*, vol. 43, no. 2, pp. 147–172, 2014.

[13] A. Flexer, F. Gouyon, S. Dixon, and G. Widmer, "Probabilistic Combination of Features for Music Classification." *ISMIR*, 2006.

[14] T. Igoe, *Groove Essentials*. Hudson Music, 2006.

[15] A. D. Sena and D. Rocchesso, "A fast Mellin and scale transform," *EURASIP Journal on Advances in Signal Processing*, 2007.

[16] W. Williams and E. Zalubas, "Helicopter transmission fault detection via time-frequency, scale and spectral methods," *Mechanical systems and signal processing*, 2000.

[17] K. Seyerlehner, M. Schedl, P. Knees, and R. Sonnleitner, "A refined block-level feature set for classification, similarity and tag prediction," *Extended Abstract to MIREX*, 2011.

[18] C. M. Bishop, *Pattern regognition and machine learning*. Springer, 2006.

[19] X. He, J. Pan, O. Jin, T. Xu, and B. Liu, "Practical Lessons from Predicting Clicks on Ads at Facebook," *ACM SIGKDD*, 2014.