USING BASS-LINE FEATURES FOR CONTENT-BASED MIR

Yusuke Tsuchihashi[†] **Tetsuro Kitahara**[†,‡] **Haruhiro Katayose**[†,‡]
[†] Kwansei Gakuin University, Japan [‡] CrestMuse Project, CREST, JST, Japan

ABSTRACT

We propose new audio features that can be extracted from bass lines. Most previous studies on content-based music information retrieval (MIR) used low-level features such as the mel-frequency cepstral coefficients and spectral centroid. Musical similarity based on these features works well to some extent but has a limit to capture fine musical characteristics. Because bass lines play important roles in both harmonic and rhythmic aspects and have a different style for each music genre, our bass-line features are expected to improve the similarity measure and classification accuracy. Furthermore, it is possible to achieve a similarity measure that enhances the bass-line characteristics by weighting the bass-line and other features. Results for applying our features to automatic genre classification and music collection visualization showed that our features improved genre classification accuracy and did achieve a similarity measure that enhances bass-line characteristics.

1 INTRODUCTION

Content-based music information retrieval (MIR) is expected to be a key technology for developing sophisticated MIR systems. One of the main issues in content-based MIR is the design of music features to be extracted from music data. The effectiveness of various features has been examined by several researchers. For audio data, Tzanetakis et al., for example, used spectral shape features (e.g., the spectral centroid, rolloff, flux, and mel-frequency cepstral coefficients (MFCCs)), rhythm content features (e.g., the beat histogram), and pitch content features [1]. Pampalk used features as zero crossing rates, mel spectral features, and fluctuation patterns (the amplitude modulation of the loudness per frequency band) [2]. Aucouturier et al. used MFCCs with various pre- and post-processing techniques [3]. Berenzweig et al. also used MFCCs [4]. Some researchers used chroma features, also known as pitch-class profiles, in addition to or instead of MFCCs [5, 6].

Most studies handling audio data have used low-level features, as described above. Low-level features such as the MFCCs and spectral centroid can be easily calculated and can capture coarse characteristics of musical content to some extent, but they have a clear limit to their ability to capture fine characteristics of musical content; it is not clear, for example, which musical aspects, such as chord voicing and instrumentation, affect the similarity of two vectors of MFCCs. The limit of low-level features was also pointed out by Pampalk [2]. Hence, trying to discover new features beyond such low-level features is a common theme in effort to improve content-based MIR.

We focus on bass lines to design new features. Bass parts play an important role for two (rhythm and harmony) of the three basic elements of music. The base lines of each music genre proceed in their own style. Moreover, a method for extracting the pitch of bass lines has been proposed [7].

In this paper, therefore, we propose new audio features that can be extracted from bass lines and describe applications of them to automatic genre classification and music collection visualization. Section 2 discusses the characteristics of bass parts and the design of audio features to be extracted from bass lines. Section 3 describes an algorithm for extracting the features. Section 3 also describes the other conventional low-level features used in the experiments. Section 4 examines the effectiveness of the proposed features in automatic genre classification tasks. Section 5 describes an application of the proposed features to music collection visualization using Music Islands [2]. We generate different views of Music Islands by switching the weights given to the features. The differences are discussed.

2 DESIGNING BASS-LINE FEATURES

2.1 Characteristics of Bass Lines

Bass lines are contained in almost all genres of music and play important roles in both harmonic and rhythmic aspects. Bass lines usually emphasize the chord tones of each chord, while they work along with the drum part and the other rhythm instruments to create a clear rhythmic pulse [8].

Bass lines have a different style for each genre. In popular music, for example, bass lines often use "riffs", which are usually simple, appealing musical motifs or phrases that are repeated, throughout the musical piece [8]. In jazz music, a sequence of equal-value (usually quarter) notes, called a *walking bass* line, with a melodic shape that alternately rises and falls in pitch over several bars, is used [9]. In rock music, a simple sequence of eighth root notes is frequently used. In dance music, a repetition of a simple phrase that sometimes involves octave pitch motions is used.

2.2 Basic Policy for Designing Bass-line Features

As described above, bass lines for each genre have a different style in both pitch and rhythm. However, we design only pitch-related features due to technical limitations. We

use the PreFEst-core [7] for extracting bass lines. It detects the pitch of the bass part, specifically the most predominant pitch in a low-pitch range, for each frame. Because it does not contain the process of sequential (temporal) grouping, the rhythm of the bass line may not be recognized.

Our pitch-related bass-line features are divided into two kinds: features of pitch variability and those of pitch motions. We design as many features as possible because we reduce the dimensionality after the feature extraction. It is known that pitch estimation techniques sometimes generate octave errors (double-pitch or half-pitch errors). We therefore extract features not only from specific pitches but also from note names (pitch classes) to avoid the influence of octave pitch-estimation errors.

2.3 Features of Pitch Variability

The following 11 features are used:

- Fv1: Number of different pitches that appear in at least one frame in the musical piece.
- Fv2: Number of pitches from Fv1 excluding those with an appearance rate of less than 10% or 20%.
- Fv3: Temporal mean of the numbers of different pitches within a sliding short-term (2 s in the current implementation) window.
- Fv4: Percentage of appearance frequencies of the i most frequently appearing pitches ($i = 1, \cdots 5$).
- Fv5: Pitch interval between the two most frequently appearing pitches.
- Fv6: Period of the most frequently appearing pitch.

Fv1 was designed to distinguish rock music, which tends to use a comparatively small number of notes in bass lines, and jazz music, which tends to use passing notes frequently. Fv2 was prepared to improve the robustness to pitch misestimation. Fv3 was designed to distinguish electronic music, which tends to repeat short phrases moving in pitch, from rock music, which tends to play root notes in bass lines. Fv4 will have high values when the same short phrase is simply repeated throughout the piece. Such repetetion is sometimes used in some types of music such as dance music. Fv5 and Fv6 were designed by referring to the pitch content features of Tzanetakis et al. [1].

2.4 Features of Pitch Motions

The following 10 features are used:

- Ft1: Mean of the numbers of pitch motions (the changes of the bass pitches) per unit time.
- Ft2: Percentage of each of the following kinds of pitch motions: chromatic, conjunct (i.e., either choromatic or diatonic), disjunct (i.e., leaps), and octave.
- Ft3: Percentage of each of the following kinds of successive pitch motions: conjunct+conjunct, conjunct+disjunct, disjunct+conjunct, and disjunct+disjunct.

Similarly to Fv1, Fv2, and Fv3, we designed Ft1 to distinguish jazz music, which tends to involve frequent pitch motions, from rock music, which tends to maintain the same note (usually the root note) within a chord. Ft2 and Ft3 were intended to distinguish walking bass lines from bass riffs involving octave motions used in electronic music.

3 ALGORITHM OF BASS-LINE FEATURE EXTRACTION

Our bass-line features and conventional low-level features are extracted through the following steps.

3.1 Extracting Pitch Trajectory of Bass line

Given an audio signal, the spectrogram is first calculated. The short-time Fourier transform shifted by 10 ms (441 points at 44.1-kHz sampling) with an 8,192-point Hamming window is used. The frequency range is then limited by using a bandpass filter to deemphasize non-bass-part features. The same filter setting as that used by Goto [7] is used.

After that, PreFEst-core [7], a multi-pitch analysis technique, is used. PreFEst-core calculates the relative predominance of a harmonic structure with every pitch in every frame. The most highly predominant pitch for each frame is basically extracted as the result of pitch estimation, but the second most highly predominant pitch is extracted if it is the same as the top one in the previous frame.

3.2 Extracting Bass-line Features

The 21 bass-line features defined in Section 2 are extracted from the pitch trajectory of the bass line. In addition, the same features are extracted from the note name trajectory (i.e., the octave is ignored) to avoid the influence of octave pitch-estimation errors; however, the percentage of octave pitch motions is excluded. The total number of bass-line features is 41.

3.3 Extracting Timbral Features

The timbral features used by Lu et al. [10] are used. The features are listed in Table 1. These features are extracted from the power spectrum in every frame and their temporal means and variances are then calculated.

3.4 Extracting Rhythmic Features

We designed rhythmic features by referring to previous studies, including that of Tzanetakis et al. [1]. For each 3-s window, auto-correlation is analyzed. The following features are then extracted:

Table 1. Timbral features used in our experiments

Intensity	Sum of intensities for all frequency bins
Sub-band intensity	Intensity of each sub-band (7 sub-bands were prepared)
Spectral centroid	Centroid of the short-time amplitude spectrum
Spectral rolloff	85th percentile of the spectral distribution
Spectral flux	2-norm distance of the frame-to-frame spectral amplitude difference)
Bandwidth	amplitude weighted average of the differences between the spectral components and the centroid
Sub-band peak	Average of the percent of the largest amplitude values in the spectrum of each sub-band
Sub-band valley	Average of the percent of the lowest amplitude values in the spectrum of each sub-band
Sub-band contrast	Difference between Peak and Valley in each sub-band

- Fr1: Ratio of the power of the highest peak to the total sum.
- Fr2: Ratio of the power of the second-highest peak to the total sum.
- Fr3: Ratio of Fr1 and Fr2.
- Fr4: Period of the first peak in BPM.
- Fr5: Period of the second peak in BPM.
- Fr6: Total sum of the power for all frames in the window.

3.5 Dimensionality Reduction

The dimensionality of the feature space is reduced to avoid the so-called curse of dimensionality. The dimensionality reduction is performed through the following three steps:

1. For every feature pair, if its correlation is higher than a threshold r, the feature that has a lower separation capacity is removed. The separation capacity is calculated as the ratio of the between-class variance to the within-class variance.
2. The s_1 features having the highest separation capacities are chosen.
3. The dimensionality is further reduced from s_1 to s_2 by using principal component analysis (PCA).

The parameters r, s_1, s_2 are determined experimentally.

4 APPLICATION TO GENRE CLASSIFICATION

In this section, we mention an application of our bass-line features to automatic genre classification. Automatic genre classification [1, 11] is a representative task in the content-based MIR field because genre labels can describe coarse characteristics of musical content despite their ambiguous definitions and boundaries. In fact, many researchers have attempted to perform automatic genre classification and entered their classification accuracies in competitions at the Music Information Retrieval Evaluation Exchange (MIREX). However, bass-line features were not used in previous audio-based genre classification studies. In this section, we show that our bass-line features are effective at improving the accuracy of automatic genre classification.

4.1 Experimental Conditions

We used an audio data set consisting of 300 musical pieces (50 for each genre) of six different genres: Pop/Rock, Metal/Punk, Electronic, Jazz/Blues, Classical, and World. This was taken from the data set distributed on the Web for the ISMIR 2004 Audio Description Contest[1]. To reduce computational costs, we used only a one-minute term for each piece. The one-minute term right after the first one minute was excerpted to avoid lead-ins, which sometimes do not contain bass lines.

Given an audio signal, the feature vector $\boldsymbol{x}$ consisting of the bass-line, timbral, and rhythmic features was extracted using the algorithm described in Section 3. The Mahalanobis distance of the feature vector $\boldsymbol{x}$ to the feature distribution of each genre c was calculated as follows:

$$D_c^2 = (\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{t}} \Sigma_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c),$$

where t is the transposition operator, and $\boldsymbol{\mu}_c$ and Σ_c are the mean vector and covariance matrix, respectively, of the feature distribution for the genre c. Finally, the genre minimizing the Mahalanobis distance, that is, $\mathrm{argmin}_c D_c^2$, was determined as the classification result.

The experiments were conducted using the leave-one-out cross validation method for each of the with- and without-bass-line conditions. The parameters r, s_1, s_2 were determined as the best parameters found through preliminary experiments for each condition: we used $r = 0.5, s_1 = 7, s_2 = 2$ for the with-bass-line condition and $r = 0.65, s_1 = 8, s_2 = 2$ for the without-bass-line condition.

4.2 Experimental Results

Experimental results are listed in Table 2. The classification accuracy was improved from 54.3% to 62.7% on average. In particular, the result for Electronic was greatly improved, from 10% to 46%. The feature distribution for each genre is shown in Figure 1. Here, the second dimension in the with-bass-line condition contributes to the separation of the distributions for Electronic and other genres. The second dimension represents the pitch variability of bass lines.

The features selected through the dimensionality reduction process for the with- and without-bass-line conditions are listed in Table 3. In the with-bass-line condition, Fv1 (the number of different pitches that appear in at least one frame) and Ft3 (the ratio of successive pitch transition patterns) were selected instead of the temporal mean of the 2nd sub-band intensity and the temporal variance of the 4th sub-band contrast, which were selected in the without-bass-line condition. The temporal mean of the 2nd sub-band intensity was extracted from a low-pitch sub-band, so it can be

[1] http://ismir2004.ismir.net/genre_contest/index.htm

Table 2. Results of automatic genre classification

(a) without bass-line features Avg: **54.3%**

	P/R	M/R	El.	J/B	Cl.	Wo.	Acc.
Pop/Rock	**21**	13	3	7	2	4	42%
Metal/Punk	6	**42**	1	1	0	0	84%
Electric	11	8	**5**	12	9	5	10%
Jazz/Blues	0	0	0	**48**	0	2	98%
Classical	0	0	0	4	**31**	15	62%
World	3	4	0	7	20	**16**	32%

(b) with bass-line features Avg: **62.7%**

	P/R	M/R	El.	J/B	Cl.	Wo.	Acc.
Pop/Rock	**23**	10	8	5	0	4	46%
Metal/Punk	5	**42**	0	2	0	1	84%
Electric	10	1	**23**	6	3	7	46%
Jazz/Blues	5	1	1	**37**	1	5	74%
Classical	0	0	0	2	**45**	3	90%
World	7	1	6	10	8	**18**	36%

Vertical axis: ground truth, horizontal axis: classification results

Table 3. Features selected with dimensionality reduction.

Without bass-line features	With bass-line features
Mean of intensity	Mean of intensity
[Mean of 2nd sub-band intensity]	Mean of 5th sub-band contrast
Mean of 5th sub-band contrast	Fv1: # of pitches that appear
Var of 6th sub-band intensity	Var of 6th sub-band intensity
[Var of 4th sub-band contrast]	Var of 6th sub-band valley
Var of 6th sub-band valley	Mean of spectral flux
	Ft3: % of disjunct+conjunct motions

[ ] denotes features selected for the without-bass-line condition.
Underline denotes selected bass-line features.

regarded as an *implicit* bass-line feature. This result shows that our *explicit* bass-line features contribute to improve separation capacity much more than such an implicit bass-line feature.

One possible reason for genre classification errors is misestimation of bass-line pitches. The misestimation tended to occur at on-beat times when drum instruments sound. One possible solution to this problem is to use drum-sound reduction techniques [12, 13].

4.3 Examination of Ground Truth by Human Subjects

Another possible reason for genre classification errors lies in ambiguous definitions and boundaries of music genres. By conducting a listening test with human subjects, we discuss the appropriateness of the ground truth for pieces whose genres were or were not successfully identified.

We first chose one piece each at random from genre-identified pieces and mis-identified pieces for each genre; the total number of chosen pieces was 12. We then asked human subjects the most likely genre for each piece. The human subjects were 10 people who often listened to music in everyday life. The results are listed in Table 4. For Pop/Rock, Metal/Punk, Electronic, and Classical, at least half the subjects chose the same genres as our system, which were different from the official ground truth. This result implies that these pieces lie on the boundaries of the genres

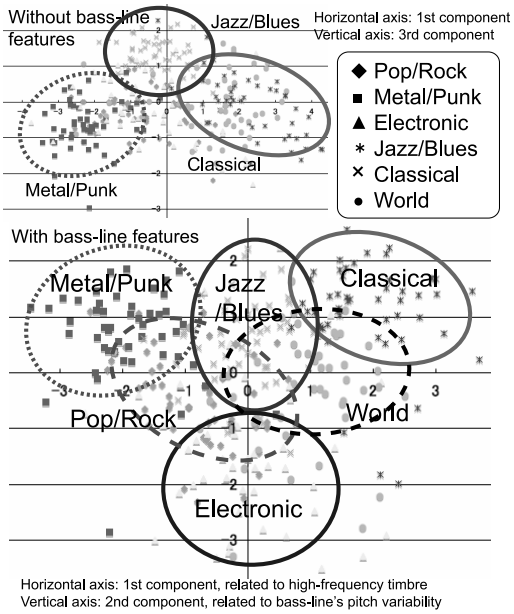


Figure 1. Feature distributions. While the distributions for Electronic and World spreaded out all over the feature space without the bass-line features, those gathered with the bass-line features.

Table 4. Results of listening test by human subjects

	P/R		M/P		El.		J/B		Cl.		Wo.	
	T	F	T	F	T	F	T	F	T	F	T	F
Pop/Rock	[7]	4	3	[**6**]	4	[**5**]		[ ]		2		1
Metal/Punk	3		[7]	3								
Electric		[**6**]		1	[4]	5		2				
Jazz/Blues					1		[10]	8				[2]
Classical									[10]	1		
World					1					[**7**]	[10]	7

T/F denotes the piece whose genre is correctly/incorrectly identified.
[ ] denotes the result of automatic genre classification.
Underline denotes the case that at least half the subjects chose the same genre as the system's output although the ground truth is different.

and that their genre classification is essentially difficult.

5 APPLICATION TO MUSIC ISLANDS

Music genres are useful concepts, but they are essentially ambiguous. Some pieces categorized into Rock/Pop could be very similar to Electronic, and others could be similar to Metal/Punk. We therefore need a mechanism for distinguishing pieces that have such different characteristics even though they are categorized into the same genre. One solution to this is to visualize music collections based on musical similarity. Because different listeners may focus on different musical aspects (e.g. melody, rhythm, timbre, and bass-line), it should be possible to adapt music similarity to such users' differences. We therefore choose to use the Music Islands technique developed by Pampalk [2] as a music collection visualizer and generate different views by switching weights given to the features.

Table 5. Settings of feature weighting

	Bass-line features	Timbral/rhythmic features
All Mix	1.0	1.0
Timbre&Rhythm	0.0	1.0
Bass Only	1.0	0.0

5.1 What are Music Islands

The key idea of Music Islands is to organize music collections on a map such that similar pieces are located close to each other. The structure is visualized using a metaphor of geographic maps. Each island represent a different style (genre) of music. This representation is obtained with an unsupervised clustering algorithm, the Self-organizing Map.

5.2 Switching Views of Music Islands

As Pampalk pointed out [2], there are various aspects of similarity. It is thus better to allow users to adjust the aspects that they are interested in exploring in their music collections. He therefore used three different aspects of features: periodicity histograms related to rhythmic characteristics, spectrum histograms related to timbre, and metadata specified manually by users. Different combinations of these aspects successfully generated different views, but there is room for improvement if features that are more clearly related to specific musical aspects are available.

In this paper, we introduce our bass-line features to generate Music Islands. By switching the weights given to the bass-line and other features, we produce music similarity measures and collection views that enhance different aspects of music.

5.3 Implementation

We implemented a system for visualizing a music collection using Matlab (Figure 2). This system generates three different views called *All Mix*, *Timbre&Rhythm*, and *Bass Only*. The settings of feature weights are listed in Table 5. The users are allowed to select one view based on their interests and to click any cell in a view to select and play back a musical piece. We used SOM Toolbox [14] and SDH Toolbox [15].

5.4 Experiments and Discussions

We conducted experiments on the generation of music collection maps to discuss the difference among three kinds of views. We used the 300 musical pieces (50 pieces per genre $\times$ 6 genres), which were the same as those used in the experiments described in the previous section. The map size was set to 14×14.

The results of generating music collection maps, shown in Figure 3, can be summarized as follows:

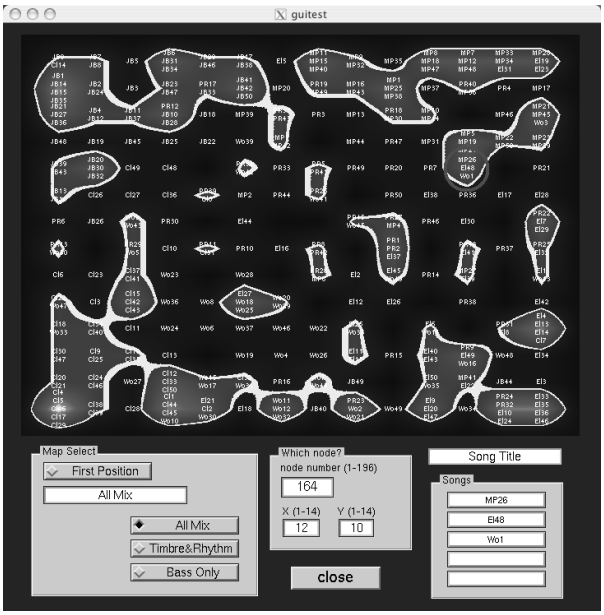
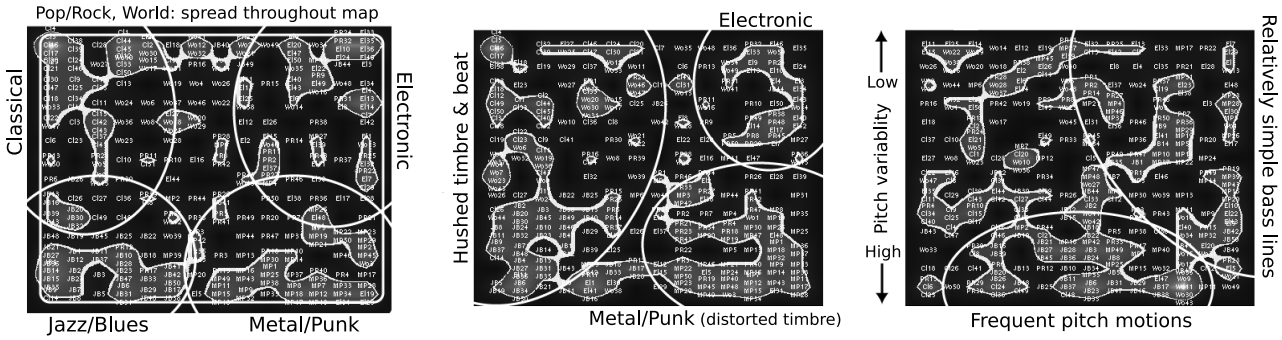


Figure 2. Prototype system of music collection visualizer.

- *All Mix*
  For Classical, Electronic, Jazz/Blues, and Metal/Punk, pieces in the same genre tended to be located close to each other. This is because these genres have comparatively clear characteristics in timbral, rhythmic, and bass-line aspects. Pop/Rock and World pieces, on the other hand, were spread throughout the map. This is because these genres cover a wide range of music, so their correspondence to acoustic characteristics is unclear. In fact, *rock* is sometimes treated as a super category of *metal* and *punk*. *Pop music* is widely used to classify any kinds of musical pieces that have a large potential audience and *world music* to classify any kinds of non-western music.
- *Timbre&Rhythm*
  The basic tendency was similar to that of *All Mix*. The major difference was that islands of Classical and Jazz/Blues were connected by a land passage. This is because music of both genres is played on similar instruments and has a comparatively weak beat. Moreover, Electronic pieces tended to spread much more throughout the map than *All Mix* ones. This result matches the improvement in genre classification.
- *Bass Only*
  Jazz/Blues pieces that had high pitch variability in their bass lines were located at the bottom of the map. On the other hand, pieces that have low pitch variability in their bass lines were located at the top of the map. Thus the map enhanced the characteristics of bass lines.

From these results, we consider that aspects enhanced in



(a) *All Mix* (b) *Timbre&Rhythm* (c) *Bass Only*

Figure 3. Music Islands with three different settings of feature weighting.

the similarity measure can be switched by changing feature weighting.

6 CONCLUSION

In this paper, we described our design of bass-line features to be extracted from audio signals and applied them to automatic genre classification and music collection visualization. Experimental results for automatic genre classification showed that the use of bass-line features improved classification accuracy from 54.3% to 62.7%. Experimental results for music collection visualization showed that we produced music collection views that could enhance different musical aspects by switching the weights to the bass-line and other features.

To achieve user-adaptive MIR, we need feature extraction techniques that can separately capture various musical aspects and feature integration techniques that are aware of users' preferences. If a user tends to focus on a specific musical aspect such as a melody, rhythm pattern, timbre, or bass line, an MIR system for this user should give higher weights to such an aspect than to other aspects. However, only a few attempts [16, 17] have been made to apply features that capture specific aspects to content-based MIR. In particular, no attempts have been made to apply bass-line features. In this paper, we described how we designed bass-line features and applied them to automatic genre classification and music collection visualization. The use of different feature weights achieved different collection views that can be switched by users according to their preferences.

Future work will include integrating our bass-line features with features capturing other musical aspects, for example, instrumentation [16], drum patterns [12], vocal timbre [17], and harmonic content [5].

7 REFERENCES

[1] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.*, 10(5):293–302, 2002.

[2] E. Pampalk. *Computational Models of Music Similarity and their Application in Music Information Retrieval*. PhD thesis, Technischen Universitat Wien, 2006.

[3] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, 2004.

[4] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whiman. A large-scale evaluation of acoustic and subjective music similarity measure. In *Proc. ISMIR*, 2003.

[5] Juan P. Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *Proc. ISMIR*, 2005.

[6] D. P. W. Ellis. Classsifying music audio with timbral and chroma features. In *Proc. ISMIR*, 2007.

[7] M. Goto. A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Comm.*, 43(4):311–329, 2004.

[8] *http://en.wikipedia.org/wiki/Bassline*.

[9] *http://en.wikipedia.org/wiki/Walking_bass*.

[10] L. Lu, D. Liu, and H.-J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(1), 2006.

[11] J.-J. Aucouturier and F. Pachet. Representing musical genres: A state of the art. *J. New Music Res.*, 32(1):83–93, 2003.

[12] K. Yoshii, M. Goto, and H. G. Okuno. Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(1):333–345, 2007.

[13] K. Miyamoto, M. Tatezono, J. L. Roux, H. Kamoeka, N. Ono, and S. Sagayama. Separation of harmonic and non-harmonic sounds based on 2d-filtering of the spectrogram. In *Proc. ASJ Autumn Meeting*, pages 825–826, 2007. (in Japanese).

[14] *http://www.cis.hut.fi/projects/somtoolbox/*.

[15] *http://www.ofai.at/~elias.pampalk/sdh/*.

[16] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Instrogram: Probabilistic representation of instrument existence for polyphonic music". *IPSJ Journal*, 48(1):214–226, 2007. (also published in IPSJ Digital Courier, Vol.3, pp.1–13).

[17] H. Fujihara and M. Goto. A music information retrieval system based on singing voice timbre. In *Proc. ISMIR*, pages 467–470, 2007.