Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004),
pp.184-191, October 2004.

AUTOMATIC DRUM SOUND DESCRIPTION FOR REAL-WORLD MUSIC USING TEMPLATE ADAPTATION AND MATCHING METHODS

Kazuyoshi Yoshii[†] *Masataka Goto*[‡] *Hiroshi G. Okuno*[†]
[†]Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University, Japan
[‡]National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

This paper presents an automatic description system of drum sounds for real-world musical audio signals. Our system can represent onset times and names of drums by means of *drum descriptors* defined in the context of MPEG-7. For their automatic description, *drum sounds must be identified* in such polyphonic signals. The problem is that acoustic features of drum sounds vary with each musical piece and precise templates for them cannot be prepared in advance. To solve this problem, we propose new template-adaptation and template-matching methods. The former method adapts a single *seed template* prepared for each kind of drums to the corresponding drum sound appearing in an actual musical piece. The latter method then can detect all the onsets of each drum by using the corresponding adapted template. The onsets of bass and snare drums in any piece can thus be identified. Experimental results showed that the accuracy of identifying bass and snare drums in popular music was about 90%. Finally, we define drum descriptors in the MPEG-7 format and demonstrate an example of the automatic drum sound description for a piece of popular music.

keywords: automatic description, polyphonic music, drum sounds, template-adaptation, template-matching

1. INTRODUCTION

The automatic description of contents of music is an important subject to realize more convenient music information retrieval. Today, audio editing, music composing and digital distribution of music are very popular because technological advances with respect to computers and the Internet are remarkable. However, we have a few efficient ways to retrieve our favorite musical pieces from huge music databases (i.e., exploration is limited to artist-based or title-based queries). In these backgrounds, many studies have addressed the content-based music information retrieval by describing music contents [4, 12, 18].

In this paper, we discuss an automatic description system of drum sounds. We aim at symbolically representing onset times and names of drums by means of *drum descriptors* defined in the context of MPEG-7. MPEG-7 is a standardization to describe contents of multimedia. Gómez *et al.* [4] and Peeters *et al.* [18] designed *instrument descriptors* in the MPEG-7 format and claimed their importance in music information retrieval. Kitahara *et al.* [12] discussed the identification of harmonic sounds to automatically describe names of instruments by using instrument descriptors. However, no research has addressed the automatic drum sound description.

Because drums play an important role in contemporary music, the drum sound description is necessary to accurately extract various features of music that are useful for music information retrieval (e.g., rhythm, tempo, beat, meter and periodicity). Previous researches, however, extracted those features by numerical analysis, not considering *symbolic information* with respect to drum performances [9, 15, 16, 20]. Some researches, for example, addressed a genre classification problem [1, 21]. Characteristic or typical drum patterns are different among genres (e.g, rock-style, jazz-style or techno-style). Therefore, symbolic information of drum sounds provides good clues for the genre classification. In addition, it distributes to music information retrieval which considers users' preferences to music because drum patterns are closely related to a mood of a musical piece [13].

It is required for the automatic drum sound description *to identify drum sounds* in real-world CD recordings. To identify instrument sounds with the harmonic structure, several methods have been proposed [2, 14]. Those methods assuming the harmonic structure, however, cannot be applied to drum sounds. Some researches addressed the drum sound identification for solo tones [8, 10, 11] or synthesized signals by MIDI [3, 5, 17]. Others discussed the extraction of drum tracks, but did not mention the identification [22]. The accurate drum sound identification for real-world polyphonic music is still difficult problem because it is impossible to prepare, in advance, all kinds of drum sounds appearing in various musical pieces.

To identify drum sounds, we propose new template adaptation and matching methods:

- The *template-adaptation method* uses template models of the power spectrum of drum sounds. The advantage of our method is that only one template model called "*seed template*" is necessary for each

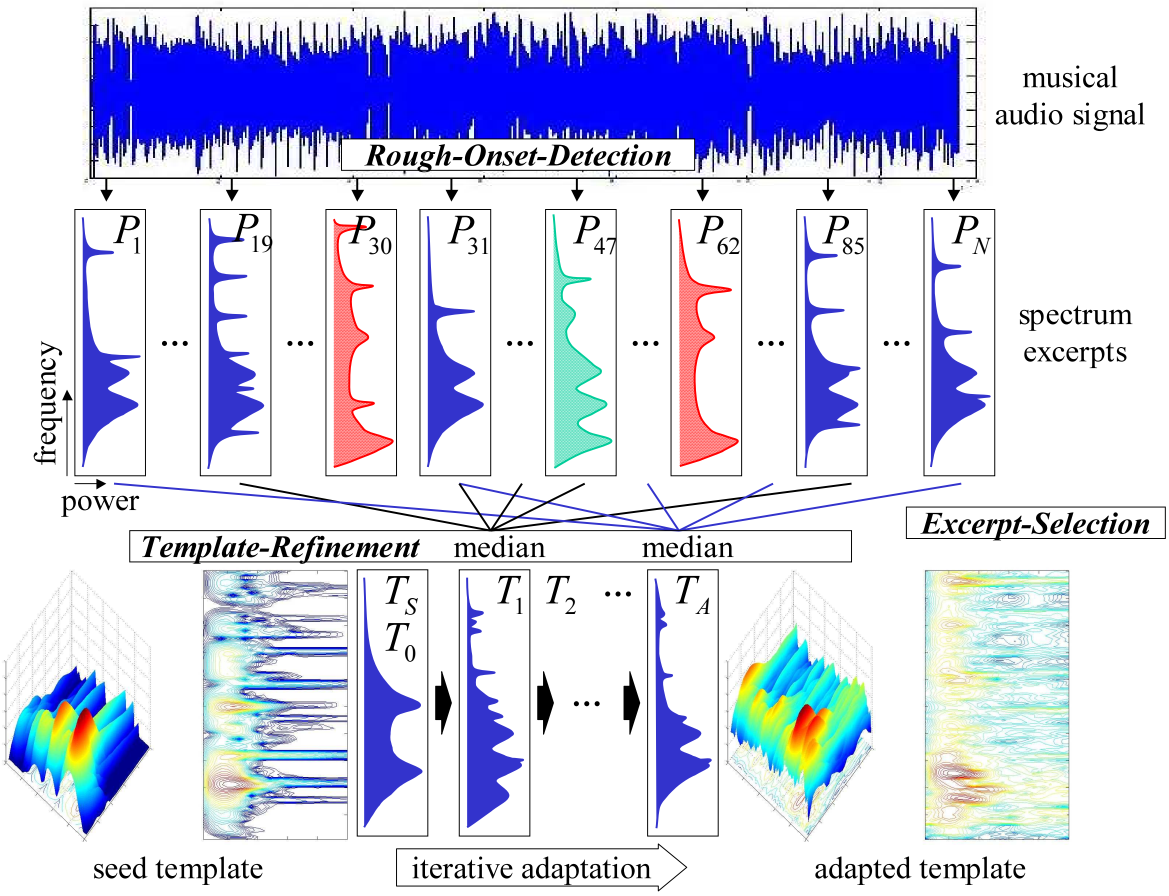


Figure 1. Overview of template-adaptation method: The template is the power spectrum in the time-frequency domain. This method adapts the single seed template to the corresponding drum sounds appearing in an actual musical piece. The method is based on an iterative adaptation algorithm, which successively applies two stages — the *Excerpt-Selection* stage and the *Template-Refinement* stage — to obtain the adapted template.

kind of drums: the method does not require a large database of drum sounds. To identify bass and snare drums, for example, we should prepare just two seed-templates (i.e., prepare a single example for each drum sound).

- The *template-matching method* is developed to identify all the onset times of drum sound after this adaptation. It uses a new distance measure that can find all the drum sounds in the piece by using the adapted templates.

The rest of this paper is organized as follows. First, Section 2 and 3 describe the template-adaptation and template-matching methods respectively to identify bass and snare drum sounds. Next, Section 4 shows experimental results of evaluating those methods. In addition, it demonstrates an example of the drum sound description by using drum descriptors defined in the standard MPEG-7 format. Finally, Section 5 summarizes this paper.

2. TEMPLATE ADAPTATION METHOD

In this paper, templates of drum sounds are the power spectrum in the time-frequency domain. The promising adaptation method of Zils *et al.* [23] worked only in the time domain because they defined templates consisting of audio signals. Extending their idea, we define templates in the time-frequency domain because non-harmonic sounds like drum sounds are well characterized by the shapes of power spectrum. Our template-adaptation method uses a single base template called "*seed template*" for each kind of drums. To identify bass and snare drums, for example, we require just two seed templates, each of which is individually adapted by the method.

Our method is based on an iterative adaptation algorithm. An overview of the method is depicted in Figure 1. First, the *Rough-Onset-Detection* stage roughly detects onset candidates in the audio signal of a musical piece. Starting from each of them, a spectrum excerpt is extracted from the power spectrum. Then, by using all the spectrum excerpts and the seed template of each drum sound, the iterative algorithm successively applies two stages — the *Excerpt-Selection* and *Template-Refinement* stages — to obtain the adapted template.

1. The *Excerpt-Selection* stage calculates the distance between the template (either the seed template or the intermediate template that is in the middle of

adaptation) and each of the spectrum excerpts by using a specially-designed distance measure. The spectrum excerpts of a certain fixed ratio to the whole are selected by ascending order with respect to the distances.

2. The *Template-Refinement* stage then updates the template by replacing it with the median of the selected excerpts. The template is thus adapted to the current piece and used for the next iteration.

Each iteration consists of these two stages and the iteration is repeated until the adapted template converges.

2.1. Rough Onset Detection

The *Rough-Onset-Detection* stage is necessary to reduce the computational cost of the two stages in the iteration. It makes it possible to extract a spectrum excerpt that starts from not every frame but every onset time. The detected rough onset times do not necessarily correspond to the actual onsets of drum sounds: they just indicate that some sounds might occur at those times.

When the power increase is high enough, the method judges that there is an onset time. Let $P(t, f)$ denote the power spectrum at frame t and frequency f and $Q(t, f)$ be the its time differential. At every frame (441 points), $P(t, f)$ is calculated by applying the STFT with Hanning windows (4096 points) to the input signal sampled at 44.1 kHz. The rough onset times are then detected as follows:

1. If $\partial P(t, f)/\partial t > 0$ is satisfied for three consecutive frames ($t = a - 1,\ a,\ a + 1$), $Q(a, f)$ is defined as
$$Q(a, f) = \left.\frac{\partial P(t, f)}{\partial t}\right|_{t=a}. \qquad (1)$$
Otherwise, $Q(a, f) = 0$.

2. At every frame *t*, the weighted summation $S(t)$ of $Q(t, f)$ is calculated by
$$S(t) = \sum_{f=1}^{2048} F(f)\, Q(t, f), \qquad (2)$$
where $F(f)$ is a function of lowpass filter that is determined as shown in Figure 2 according to the frequency characteristics of typical bass or snare drum sounds.

3. Each onset time is given by the peak time found by peak-picking in $S(t)$. $S(t)$ is smoothed by the Savitzky and Golay's smoothing method [19] before its peak time is calculated.

2.2. Seed Template and Spectrum Excerpt Preparation

Seed template T_S, which is a spectrum excerpt prepared for each of bass and snare drums, is created from audio signal of an example of that drum sound, which must be monophonic (solo tone). By applying the same method with the *Rough-Onset-Detection* stage, an onset time in the audio signal is detected. Starting from the onset time, T_S is extracted from the STFT power spectrum of the signal. T_S is represented as a time-frequency matrix whose element is denoted as $T_S(t, f)$ ($1 \le t \le 15$ [frames], $1 \le f \le 2048$ [bins]). In the iterative adaptation algorithm, a template being adapted after g-th iterations is denoted as T_g. Because T_S is the first template, T_0 is set to T_S.

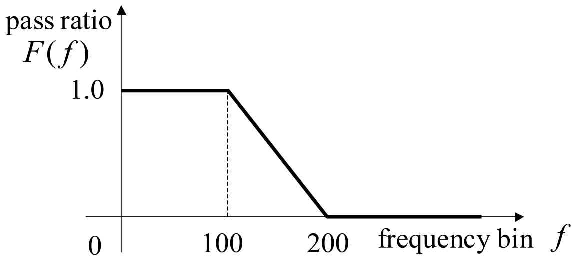


Figure 2. Function of the lowpass filter according to the frequency characteristics of typical bass and snare drums.

On the other hand, spectrum excerpt P_i is extracted starting from each detected onset time o_i ($i = 1, \cdots, N$) [ms] in the current musical piece. N is the number of the detected onsets in the piece. The spectrum excerpt P_i is also represented as a time-frequency matrix whose size is same with the template T_g.

We also obtain $\acute{T}_g$ and $\acute{P}_i$ from the power spectrum weighted by the lowpass filter $F(f)$:
$$\acute{T}_g(t, f) = F(f)\, T_g(t, f), \qquad (3)$$
$$\acute{P}_i(t, f) = F(f)\, P_i(t, f). \qquad (4)$$

Because the time resolution of the onset times roughly estimated is 10 [ms] (441 points), it is not enough to obtain high-quality adapted templates. We therefore adjust each rough onset time o_i [ms] to obtain more accurate spectrum excerpt P_i extracted from adjusted onset time o'_i [ms]. If the spectrum excerpt from $o_i - 5$ [ms] or $o_i + 5$ [ms] is better than that from o_i [ms], o'_i [ms] is set to the time providing the better spectrum excerpt as follows:

1. The following is calculated for $j = -5, 0, 5$.
   (a) Let $P_{i,j}$ be a spectrum excerpt extracted from $o_i + j$ [ms]. Note that the STFT power spectrum should be calculated again for $o_i + j$ [ms].
   (b) The correlation *Corr*(j) between the template T_g and the excerpt $P_{i,j}$ is calculated as
$$\mathit{Corr}(j) = \sum_{t=1}^{15} \sum_{f=1}^{2048} \acute{T}_g(t, f)\, \acute{P}_{i,j}(t, f), \qquad (5)$$
where $\acute{P}_{i,j}(t, f) = F(f)\, P_{i,j}(t, f)$.

2. The best index J is determined as index j that maximizes *Corr*(j).
$$J = \underset{j}{\operatorname{argmax}}\, \mathit{Corr}(j). \qquad (6)$$

3. P_i is determined as $P_{i,J}$.

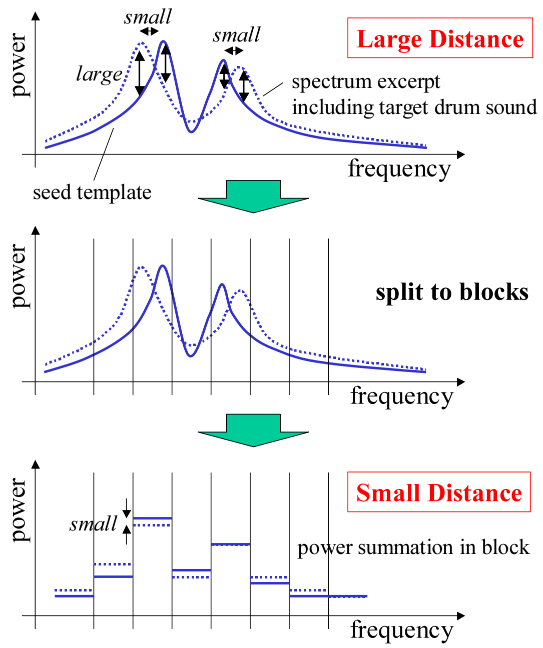


Figure 3. Our improved log-spectral distance measure to calculate the appropriate distance (quantization at a lower frequency resolution).

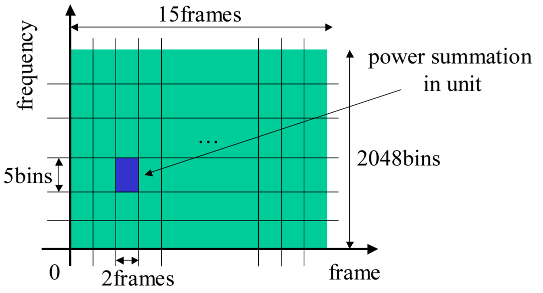


Figure 4. Our implementation of the quantization at a lower time-frequency resolution for our improved log-spectral distance measure.

2.3. Excerpt Selection

To select a set of spectrum excerpts that are similar to the intermediate template T_g, we propose an *improved log-spectral distance measure* as shown in Figure 3. The spectrum excerpts whose distance from the template is smaller than a threshold are selected. The threshold is determined so that the ratio of the number of selected excerpts to the total number is a certain value. We cannot use a normal log-spectral distance measure because it is too sensitive to the difference of spectral peak positions. Our *improved log-spectral distance measure* uses two kinds of the distance D_i — D_i for the first iteration ($g = 0$) and D_i for the other iterations ($g \geq 1$) — to robustly calculate the appropriate distance even if frequency components of the same drum may vary during a piece.

The distance D_i for the first iteration are calculated after quantizing T_g and P_i at a lower time-frequency resolution. As is shown in Figure 4, the time and frequency resolution after the quantization is 2 [frames] (20 [ms]) and 5 [bins] (54 [Hz]), respectively. The distance D_i between $T_g(T_S)$ and P_i is defined as

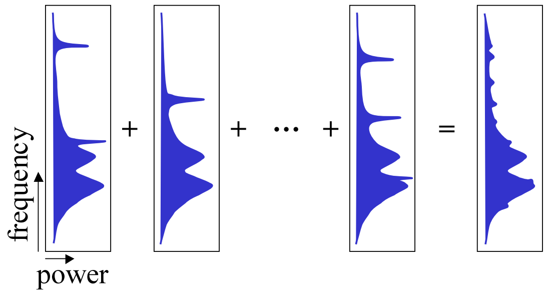


Figure 5. Updating the template by calculating the median of selected spectrum excerpts.

$$D_i = \sqrt{\sum_{\hat{t}=1}^{15/2} \sum_{\hat{f}=1}^{2048/5} \left(\hat{T}_g(\hat{t},\hat{f}) - \hat{P}_i(\hat{t},\hat{f})\right)^2} \quad (g=0), \quad (7)$$

where the quantized (smoothed) spectrum $\hat{T}_g(\hat{t},\hat{f})$ and $\hat{P}_i(\hat{t},\hat{f})$ are defined as

$$\hat{T}_g(\hat{t},\hat{f}) = \sum_{t=2\hat{t}-1}^{2\hat{t}} \sum_{f=5\hat{f}-4}^{5\hat{f}} \acute{T}_g(t,f), \quad (8)$$

$$\hat{P}_i(\hat{t},\hat{f}) = \sum_{t=2\hat{t}-1}^{2\hat{t}} \sum_{f=5\hat{f}-4}^{5\hat{f}} \acute{P}_i(t,f). \quad (9)$$

On the other hand, the distance D_i for the iterations after the first iteration is calculated by the following normal log-spectral distance measure:

$$D_i = \sqrt{\sum_{t=1}^{15} \sum_{f=1}^{2048} \left(\acute{T}_g(t,f) - \acute{P}_i(t,f)\right)^2} \quad (g \geq 1). \quad (10)$$

2.4. Template Refinement

As is shown in Figure 5, the median of all the selected spectrum excerpts is calculated and the updated (refined) template T_{g+1} is obtained by

$$T_{g+1}(t,f) = \underset{s}{\text{median}}\, P_s(t,f), \quad (11)$$

where P_s ($s = 1, \cdots, M$) are spectrum excerpts selected in the *Excerpt-Selection* stage.

We use the median operation because it can suppress frequency components that do not belong to drum sounds. Since major original frequency components of a target drum sound can be expected to appear at the same positions in most selected spectrum excerpts, they are preserved after the median operation. On the other hand, frequency components of other musical instrument sounds do not always appear at similar positions in the selected spectrum excerpts. When the median is calculated at t and f, those unnecessary frequency components become outliers and can be suppressed. We can thus obtain the drum-sound template adapted to the current musical piece even if it contains simultaneous sounds of various instruments.

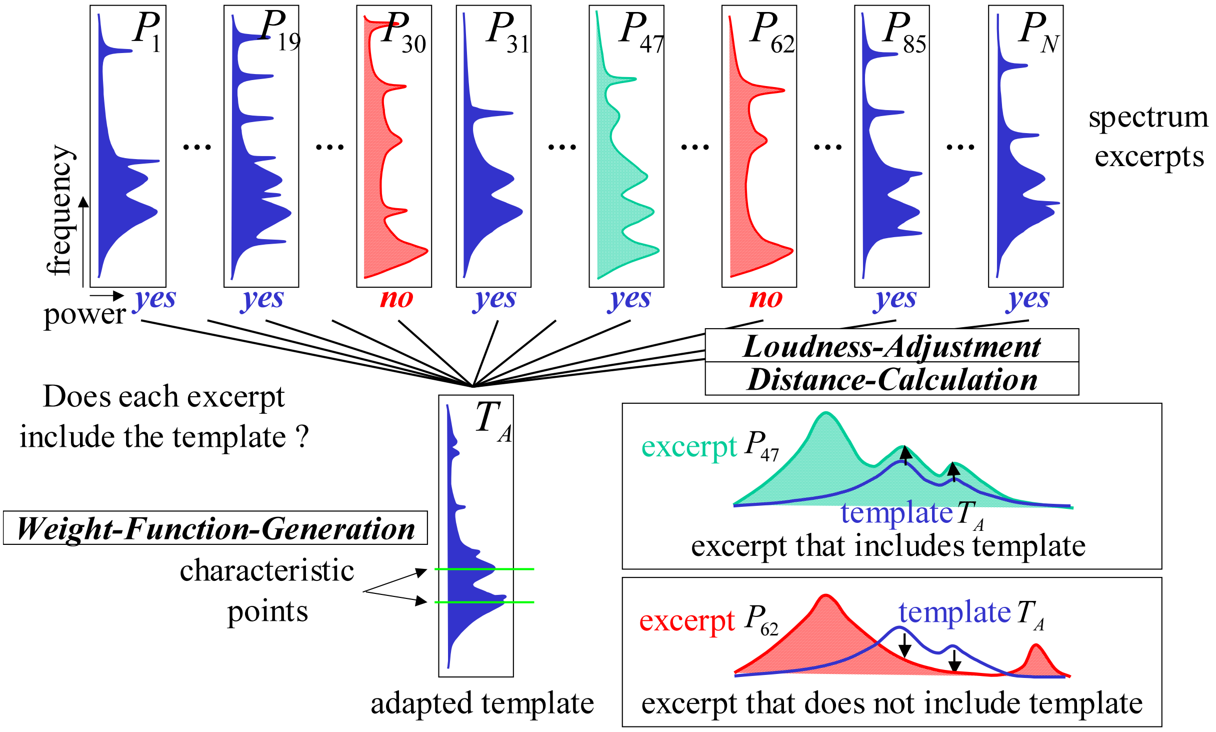


Figure 6. Overview of template-matching method: This method matches the adapted template with all spectrum excerpts by using the improved Goto's distance measure to detect all the actual onset times. Our distance measure can judge whether the adapted template is included in spectrum excerpts even if there are other simultaneous sounds.

3. TEMPLATE MATCHING METHOD

By using the template adapted to the current musical piece, this method finds all temporal locations where a targeted drum occurs in the piece: it tries to exhaustively find all onset times of the target drum sound. This template-matching problem is difficult because sounds of other musical instruments often overlap the drum sounds corresponding to the adapted template. Even if the target drum sound is included in a spectrum excerpt, the distance between the adapted template and the excerpt becomes large when using most typical distance measures. To solve this problem, we propose a new distance measure that is based on the distance measure proposed by Goto and Muraoka [5]. Our distance measure can judge whether the adapted template is included in spectrum excerpts even if there are other simultaneous sounds. This judgment is based on characteristic points of the adapted template in the time-frequency domain.

An overview of our method is depicted in Figure 6. First, the *Weight-Function-Generation* stage prepares a weight function which represents spectral characteristic points of the adapted template. Next, the *Loudness-Adjustment* stage calculates the loudness difference between the template and each spectrum excerpt by using the weight function. If the loudness difference is larger than a threshold, it judges that the target drum sound does not appear in that excerpt, and does not execute the subsequent processing. If the difference is not too large, the loudness of each spectrum excerpt is adjusted to compensate for the loudness difference. Finally, the *Distance-Calculation* stage calculates the distance between the adapted template and each adjusted spectrum excerpt. If the distance is smaller than a threshold, it judges that that excerpt includes the target drum sound.

3.1. Weight Function Generation

A weight function represents the magnitude of spectral characteristic at each frame t and frequency f in the adapted template. The weight function w is defined as

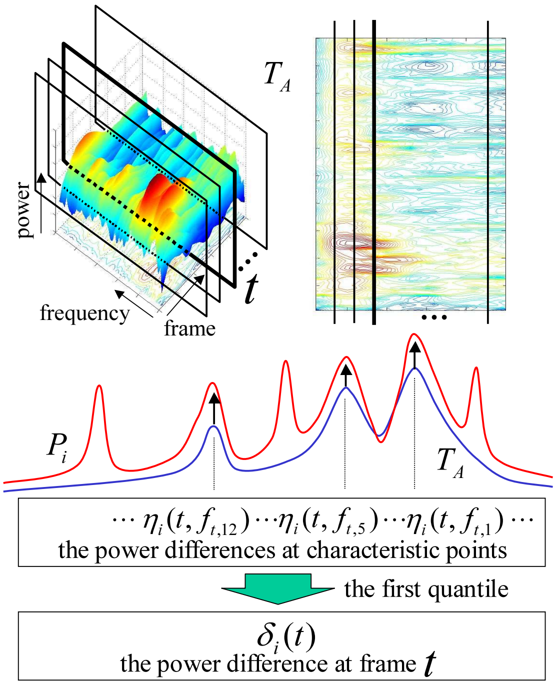
$$w(t,f) = F(f)\, T_A(t,f), \qquad (12)$$

where T_A is the adapted template and $F(f)$ is the low-pass filter function depicted in Figure 2.

3.2. Loudness Adjustment

The loudness of each spectrum excerpt is adjusted to that of the adapted template T_A. This is required by our template-matching method: if the loudness is different, our method cannot estimate the appropriate distance between a spectrum excerpt and the adapted template because it cannot judge whether the spectrum excerpt includes the adapted template.

To calculate the loudness difference between the spectrum excerpt P_i and the template T_A, we focus on spectral characteristic points of T_A in the time-frequency domain. First, spectral characteristic points (frequencies) at each frame are determined by using the weight function w, and the power difference η_i at each spectral characteristic



T_A
power
frequency
frame
t
P_i
T_A
$\cdots \eta_i(t, f_{t,12}) \cdots \eta_i(t, f_{t,5}) \cdots \eta_i(t, f_{t,1}) \cdots$
the power differences at characteristic points
the first quantile
$\delta_i(t)$
the power difference at frame t

Figure 7. Calculating the power difference $\delta_i(t)$ at each frame t, determined as the first quantile of $\eta_i(t, f_{t,k})$.

point is calculated. Next, the power difference δ_i at each frame is calculated by using η_i at that frame, as is shown in Figure 7. If the power of P_i is too much smaller than that of T_A, the method judges that P_i does not include T_A, and does not proceed with the following processing. Finally, the total power difference Δ_i is calculated by integrating δ_i. The algorithm is described as follows:

1. Let $f_{t,k}$ $(k = 1, \cdots, 15)$ be the characteristic points of the adapted template. $f_{t,k}$ represents a frequency where $w(t, f_{t,k})$ is the k-th largest at frame t. The power difference $\eta_i(t, f_{t,k})$ is calculated as

$$\eta_i(t, f_{t,k}) = P_i(t, f_{t,k}) - T_A(t, f_{t,k}). \qquad (13)$$

2. The power difference $\delta_i(t)$ at frame t is determined as the first quantile of $\eta_i(t, f_{t,k})$.

$$\delta_i(t) = \underset{k}{\text{first-quantile}}\ \eta_i(t, f_{t,k}), \qquad (14)$$

$$K_i(t) = \underset{k}{\text{arg first-quantile}}\ \eta_i(t, f_{t,k}). \qquad (15)$$

If the number of frames where $\delta_i(t) \leq \Psi$ is satisfied is larger than threshold R_δ, we judge that T_A is not included in P_i (Ψ is a negative constant).

3. The total power difference Δ_i is calculated as

$$\Delta_i = \frac{\sum_{\{t|\delta_i(t)>\Psi\}} \delta_i(t)\ w(t, f_{t,K_i(t)})}{\sum_{\{t|\delta_i(t)>\Psi\}} w(t, f_{t,K_i(t)})}. \qquad (16)$$

If $\Delta_i \leq \Theta_\Delta$ is satisfied, we judge that T_A is not included in P_i (Θ_Δ is a threshold). Let P_i' be an adjusted spectrum excerpt after the loudness adjustment, determined as

$$P_i'(t, f) = P_i(t, f) - \Delta_i. \qquad (17)$$

3.3. Distance Calculation

The distance between the adapted template T_A and the adjusted spectrum excerpt P_i' is calculated by using an extended version of the Goto's distance measure [5]. If $P_i'(t, f)$ is larger than $T_A(t, f)$ — i.e., $P_i'(t, f)$ includes $T_A(t, f)$, $P_i'(t, f)$ can be considered a mixture of frequency components of not only the targeted drum but also other musical instruments. We thus define the distance measure as

$$\gamma_i(t, f) = \begin{cases} 0 & (P_i'(t, f) - T_A(t, f) \geq \Psi), \\ 1 & \text{otherwise}, \end{cases} \qquad (18)$$

where $\gamma_i(t, f)$ is the local distance between T_A and P_i' at t and f. The negative constant Ψ makes this distance measure robust for the small variation of frequency components. If $P_i'(t, f)$ is larger than about $T_A(t, f)$, $\gamma_i(t, f)$ becomes zero.

The total distance Γ_i is calculated by integrating γ_i in the time-frequency domain, weighted by the function w:

$$\Gamma_i = \sum_{t=1}^{15} \sum_{f=1}^{2048} w(t, f)\ \gamma_i(t, f). \qquad (19)$$

To determine whether the targeted drum played at P_i', distance Γ_i is compared with threshold Θ_Γ. If $\Gamma_i < \Theta_\Gamma$ is satisfied, we judge that the targeted drum played.

4. EXPERIMENTS AND RESULTS

Drum sound identification for polyphonic musical audio signals was performed to evaluate the accuracy of identifying bass and snare drums by our proposed method. In addition, we demonstrate an example of the drum sound description by means of *drum descriptors* in MPEG-7.

4.1. Experimental Conditions

We tested our method on excerpts of ten songs included in the popular music database *RWC-MDB-P-2001* developed by Goto *et al.* [6]. Each excerpt was taken from the first minute of a song. The songs we used included sounds of vocals and various instruments as songs in commercial CDs do. Seed templates were created from solo tones included in the musical instrument sound database *RWC-MDB-I-2001* [7]: the seed templates of bass and snare drums are created from sound files named 421BD1N3.WAV and 422SD5N3.WAV respectively. All data were sampled at 44.1 kHz with 16 bits.

We evaluated the experimental results by the recall rate, the precision rate and the F-measure:

$$\text{recall rate} = \frac{\text{the number of correctly detected onsets}}{\text{the number of actual onsets}},$$

$$\text{precision rate} = \frac{\text{the number of correctly detected onsets}}{\text{the number of onsets detected by matching}},$$

$$\text{F-measure} = \frac{2 \cdot \text{recall rate} \cdot \text{precision rate}}{\text{recall rate} + \text{precision rate}}.$$

To prepare actual onset times (correct answers), we extracted onset times of bass and snare drums from the standard MIDI file of each piece, and adjusted them to the piece by hands.

piece		bass drum			snare drum		
number	method	recall rate	precision rate	F-measure	recall rate	precision rate	F-measure
No. 6	base	26 % (28/110)	68 % (28/41)	0.37	83 % (52/63)	83 % (52/61)	0.83
	adapt	57 % (63/110)	84 % (63/75)	0.68	100 % (63/63)	97 % (63/65)	0.98
No. 11	base	54 % (28/52)	100 % (28/28)	0.70	27 % (10/37)	71 % (10/14)	0.33
	adapt	100 % (52/52)	100 % (52/52)	1.00	95 % (35/37)	92 % (35/38)	0.93
No. 18	base	26 % (35/134)	100 % (35/35)	0.41	91 % (122/134)	82 % (122/148)	0.86
	adapt	97 % (130/134)	71 % (130/183)	0.82	76 % (102/134)	94 % (102/109)	0.84
No. 20	base	95 % (60/63)	100 % (60/60)	0.98	24 % (15/63)	94 % (15/16)	0.38
	adapt	94 % (59/63)	100 % (59/59)	0.97	78 % (49/63)	91 % (49/54)	0.84
No. 30	base	19 % (25/130)	89 % (25/28)	0.31	27 % (19/70)	90 % (19/21)	0.42
	adapt	93 % (121/130)	94 % (121/129)	0.93	100 % (70/70)	96 % (70/73)	0.98
No. 44	base	6 % (6/99)	100 % (6/6)	0.11	9 % (7/80)	88 % (7/8)	0.16
	adapt	93 % (92/99)	100 % (92/92)	0.96	68 % (54/80)	89 % (54/61)	0.77
No. 47	base	77 % (46/60)	98 % (46/47)	0.86	41 % (21/51)	70 % (21/30)	0.52
	adapt	93 % (56/60)	98 % (56/57)	0.96	88 % (45/51)	75 % (45/60)	0.81
No. 50	base	92 % (61/66)	94 % (61/65)	0.93	94 % (102/108)	89 % (102/114)	0.92
	adapt	97 % (64/66)	88 % (64/73)	0.92	67 % (72/108)	96 % (72/77)	0.78
No. 52	base	86 % (113/131)	96 % (113/118)	0.90	97 % (76/78)	94 % (76/81)	0.96
	adapt	94 % (123/131)	90 % (123/136)	0.92	90 % (70/78)	97 % (70/72)	0.93
No. 61	base	96 % (73/76)	100 % (73/73)	0.98	99 % (66/67)	80 % (66/83)	0.88
	adapt	93 % (71/76)	100 % (71/71)	0.97	99 % (66/67)	100 % (66/66)	0.99
average	base	51.6 % (475/951)	94.8 % (475/501)	0.67	65.2 % (490/751)	84.6 % (490/579)	0.74
	adapt	90.2 % (831/921)	90.0 % (831/927)	0.90	83.4 % (626/751)	92.7 % (626/675)	0.88

Table 1. Experimental results of drum sound identification for ten musical pieces in RWC-MDB-P-2001.

identified drum	R_δ	Ψ	Θ_Δ	Θ_Γ
(*method*)	[frames]	[dB]	[dB]	
bass drum (*base*)	7	-10	1	5000
bass drum (*adapt*)	7	-10	-10	5000
snare drum (*base*)	7	-10	-5	5000
snare drum (*adapt*)	7	-10	-7	5000

Table 2. Thresholds used in four experimental settings.

4.2. Results of Drum Sound Identification

Table 1 shows the experimental results of comparing our template-adaptation-and-matching methods (called *adapt method*) with a method in which the template-adaptation method was disabled (called *base method*); the base method used a seed template instead of the adapted one for the template matching. In other words, we conducted four experiments in different settings; the identification of bass drum by the *base* or *adapt* method and that of snare drum by the *base* or *adapt* method. We used different thresholds shown in Table 2 among four experimental cases to product the best results in respective case.

These results showed the effectiveness of the *adapt* method: the template-adaptation method improved the F-measure of identifying bass drum from 0.67 to 0.90 and that of identifying snare drum from 0.74 to 0.88 on average of the ten pieces. In fact, in our observation, the template-adaptation method absorbed the difference of the timber by correctly adapting seed templates to actual drum sounds appearing in a piece.

In many musical pieces, the recall rate was significantly improved in the *adapt* method. The *base* method often detected a few onsets in some piece (e.g., No. 11 and No. 30) because the distance between an unadapted seed template and spectrum excerpts were not appropriate; the distance became too large because of the difference of the timber. On the other hand, the template-matching method of the *adapt* method worked effectively; all the rates in No. 11 and No. 30, for example, were over 90% in the *adapt* method. If the difference of the timber is small, the *base* method produced the high recall and precision rates (e.g., No. 52 and No. 61).

Although our *adapt* method is effective in general, it caused a low recall rate in a few cases. The recall rate of identifying the snare drum in No. 50, for example, was degraded, while the precision rate was improved. In this piece, the template-matching method was not able to judge that the template was correctly included in spectrum excerpts because frequency components of the bass guitar often overlapped spectral characteristic points of the bass drum in those excerpts.

4.3. Demonstration of Drum Sound Description

In this section, we demonstrate an example of the automatic drum sound description by using *drum descriptors*. Our proposed template-adaptation and template-matching methods can detect onset times of bass and snare drums respectively. To symbolically represent these information in the context of MPEG-7, drum descriptors and their schemes must be defined in the MPEG-7 format.

First, we define drum descriptors and drum descriptor schemes. To describe onset times and names of drums, we use the *mpeg7:MediaTimePoint* data type and the *Enumeration* facet respectively:

```
<simpleType name="InstrumentNameType">
  <restriction base="string">
    <enumeration value="BassDrum"/>
    <enumeration value="SnareDrum"/>
    ...
  </restriction>
</simpleType>
<complexType name="InstrumentOnsetType">
  <sequence>
    <element name="MediaTimePoint"
      type="mpeg7:MediaTimePointType"/>
    <element name="InstrumentName"
      type="InstrumentNameType"/>
  </sequence>
</complexType>
<complexType name="InstrumentStreamType">
  <sequence>
    <element name="InstrumentOnset"
      minOccurs="0" maxOccurs="unbounded"/>
  </sequence>
</complexType>
```

where the *InstrumentOnsetType* data type indicates information of a time and a name which corresponds to a onset in a musical piece. The *InstrumentStreamType* data type is a set of multiple *InstrumentOnsetType* elements.

Next, we describe onset times and names of drums in a musical piece by means of drum descriptors defined above. We demonstrate an example of the drum sound description for No. 52 by using our proposed methods.

```
<element name="DrumStream" type="InstrumentStreamType"/>
<DrumStream>
  <InstrumentOnset>
    <MediaTimePoint>T00:00:36382F44100</MediaTimePoint>
    <InstrumentName>BassDrum</InstrumentName>
  </InstrumentOnset>
  <InstrumentOnset>
    <MediaTimePoint>T00:00:54684F44100</MediaTimePoint>
    <InstrumentName>SnareDrum</InstrumentName>
  </InstrumentOnset>
  <InstrumentOnset>
    <MediaTimePoint>T00:01:22506F44100</MediaTimePoint>
    <InstrumentName>BassDrum</InstrumentName>
  </InstrumentOnset>
  ...
</DrumStream>
```

5. CONCLUSION

In this paper, we have presented an automatic description system that can describe onset times and names of drums by means of *drum descriptors*. Our system used two methods to identify all the onset times of bass and snare drums respectively in real-world CD recordings. Even if drum sounds prepared as *seed templates* are different from ones used in a musical piece, our template-adaption method can adapt the templates to the piece. By using the adapted templates, our template-matching method then detects all the onset times. Our experimental results have shown that the adaptation method largely improved the F-measure of identifying bass and snare drums. In addition, we defined drum descriptors in the context of MPEG-7 and demonstrated the automatic drum sound description for a real-world musical piece. In the future, we plan to use multiple seed templates for each kind of drums and extend our method to identify other drum sounds.

Acknowledgments

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research (A), No.15200015, and COE program of MEXT, Japan.

6. REFERENCES

[1] Dixon, S., Pampalk, E., and G. Widmer, G., “Classification of Dance Music by Periodicity Patterns,” *Proc. of ISMIR*, 159–165, 2003.

[2] Eronen, A. and Klapuri, A., “Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features,” *Proc. of ICASSP*, 753–756, 2000.

[3] FitzGerald, D., Coyle, E., and Lawlor, B., “Sub-band Independent Subspace Analysis for Drum Transcription,” *Proc. of DAFX*, 65–69, 2002.

[4] Gómez, E., Gouyon, F., Herrera, P., and Amatriain, X., “Using and enhancing the current MPEG-7 standard for a music content processing tool,” *Proc. of AES*, 2003.

[5] Goto, M. and Muraoka, Y., “A Sound Source Separation System for Percussion Instruments,” *IEICE Transactions*, J77-D-II, 5, 901–911, 1994 *(in Japanese)*.

[6] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R., “RWC Music Database: Popular, Classical, and Jazz Music Databases,” *Proc. of ISMIR*, 287–288, 2002.

[7] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R., “RWC Music Database: Music Genre Database and Musical Instrument Sound Database,” *Proc. of ISMIR*, 229–230, 2003.

[8] Gouyon, F. and Herrera, P., “Exploration of techniques for automatic labeling of audio drum tracks instruments,” *Proc. of AES*, 2001.

[9] Gouyon, F. and Herrera, P., “Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors,” *Proc. of AES*, 2003.

[10] Herrera, P., Yeterian, A., and Gouyon, F., “Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques,” *Proc. of ICMAI*, LNAI2445, 69–80, 2002.

[11] Herrera, P., Dehamel, A., and Gouyon, F., “Automatic labeling of unpitched percussion sounds,” *Proc. of AES*, 2003.

[12] Kitahara, T., Goto, M., and Okuno, H.G., ”Category-level Identification of Non-registered Musical Instrument Sounds,” *Proc. of ICASSP*, 2004 *(in press)*.

[13] Liu, D., Lu, L., and Zhang, H.J., “Automatic Mood Detection from Acoustic Music Data,” *Proc. of ISMIR*, 2003.

[14] Martin, K.D., “Musical Instrumental Identification: A Pattern-Recognition Approach,” *136th meeting of American Statistical Association*, 1998.

[15] Pampalk, E., Dixon, S., and Widmer, G., “Exploring Music Collections by Browsing Different Views,” *Proc. of ISMIR*, 201–208, 2003.

[16] Paulus, J. and Klapuri, A., “Measuring the Similarity of Rhythmic Patterns,” *Proc. of ISMIR*, 150–156, 2002.

[17] Paulus, J. and Klapuri, A., “Model-based Event Labeling in the Transcription of Percussive Audio Signals,” *Proc. of DAFX*, 1–5, 2003.

[18] Peeters, G., McAdams, S., and Herrera, P., “Instrument Sound Description in the Context of MPEG-7,” *Proc. of ICMC*, 2000.

[19] Savitzky, A. and Golay, M., “Smoothing and Differentiation of Data by Simplified Least Squares Procedures,” *J. of Analytical Chemistry*, 36, 8, 1627–1639, 1964.

[20] Scheirer, E.D., “Tempo and Beat Analysis of Acoustic Musical Signals,” *J. of Acoustical Society of America*, 103, 1, 588–601, 1998.

[21] Tzanetakis, G. and Cook, P., “Musical Genre Classification of Audio Signals,” *IEEE Transactions on Speech and Audio Processing*, 10, 5, 2002.

[22] Uhle, C., Dittmar, C., and Sporer, T., “Extraction of Drum Tracks from Polyphonic Music Using Independent Subspace Analysis,” *Proc. of ICA*, 843–848, 2003.

[23] Zils, A., Pachet, F., Delerue, O., and Gouyon, F., “Automatic Extraction of Drum Tracks from Polyphonic Music Signals,” *Proc. of WEDELMUSIC*, 179–183, 2002.