Drum Sound Identification for Polyphonic Music Using Template Adaptation and Matching Methods

Kazuyoshi Yoshii,[†] *Masataka Goto*[‡] and *Hiroshi G. Okuno*[†]

[†]Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University, Japan
yoshii@kuis.kyoto-u.ac.jp

[‡]National Institute of Advanced Industrial
Science and Technology (AIST), Japan
m.goto@aist.go.jp

okuno@i.kyoto-u.ac.jp

Abstract

This paper describes drum sound identification for polyphonic musical audio signals. It is difficult to identify drum sounds in such signals because acoustic features of those 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 containing sounds of various musical instruments. The latter method then uses a carefully-designed distance measure that can detect all the onset times of each drum in the same piece by using the corresponding adapted template. The onset times of bass and snare drums in any piece can thus be identified even if their timbres are different from prepared templates. Experimental results with our methods showed that the accuracy of identifying bass and snare drums in popular music was about 90%.

1. Introduction

Musical instrument identification as well as automatic music transcription become important to archive and retrieve a deluge of musical audio signals. If the names of musical instruments in musical pieces can automatically be identified, they are useful for classifying music and indexing music structure. To identify musical instrument sounds with the harmonic structure, several methods have been proposed. Martin *et al.* [7] and Eronen *et al.* [1], for example, discussed identification of solo tones. Kashino *et al.* [6] developed an automatic transcription system that can identify sound sources for polyphonic music.

Because those previous methods assuming the harmonic structure cannot be applied to drum sounds, different approaches have been proposed for drum sounds. Herrera *et al.* [5] used a method of using spectral and temporal features of drum sounds and achieved the accuracy of about 90% on 643 solo tones of drum sounds. This method, however, cannot be applied to polyphonic musical audio signals including drum sounds. On the other hand, Zils *et al.* [8] proposed a time-domain method of extracting drum sounds from such polyphonic signals. They show the effectiveness of a promising idea of adapting simple templates of drum sounds to a musical piece in the time domain. This method, however, focused on resynthesizing high-quality drum sounds and did not aim at identifying all the onset times of drum sounds in a piece. The accurate identification of drum sounds in real-world polyphonic musical audio signals is still difficult problem because it is impossible to prepare, in advance, all kinds of drum sounds appearing in various musical pieces.

In this paper, we propose a frequency-domain template-adaptation method that uses the power spectrum of drum sounds as template models. The advantage of our method is that only one template model called "*seed template*" is necessary for each 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). Given the seed templates, our method can adapt them to actual drum sounds appearing in any polyphonic musical piece that contains other musical instrument sounds. To identify all the onset times of drum sounds after this adaptation, we then developed another method for accurate template-matching. 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 method and the template-matching method, respectively. Next, Section 4 shows experimental results of evaluating those methods. 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 adaptation method of Zils *et al.* [8] 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.

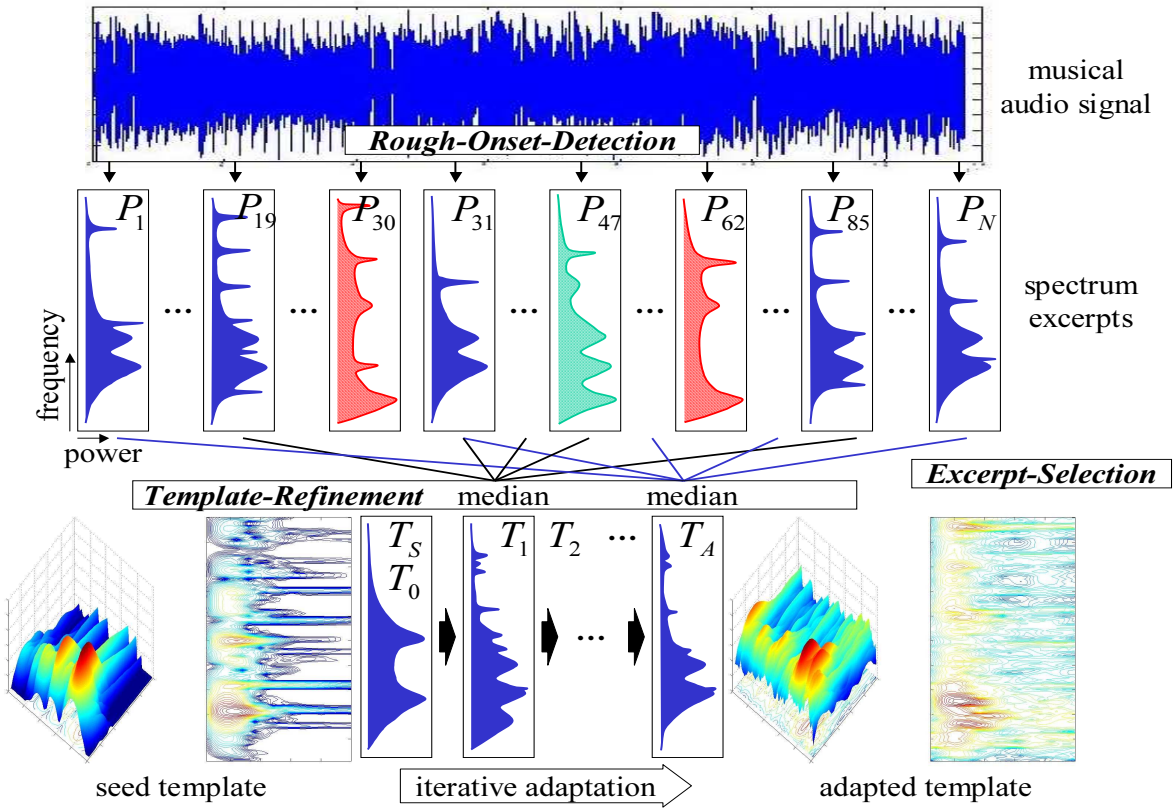

Figure 1: *Overview of template-adaptation method (iterative adaptation algorithm).*

Our method is based on an iterative adaptation algorithm. An overview of the method is depicted in Fig. 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 excepts and the seed template of each kind of drums, the iterative algorithm successively applies two stages — the *Excerpt-Selection* stage and the *Template-Refinement* stage — to obtain the adapted template.

In each iteration, the *Excerpt-Selection* stage calculates the distance between the template (the seed template is used for the first iteration) and each of the spectrum excerpts by using a specially-designed distance measure. It selects a set of spectrum excerpts whose distance is smaller (the ratio of the set to the whole is a constant). 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. 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
$$Q(a, f) = \left.\frac{\partial P(t, f)}{\partial t}\right|_{t=a}.$$
Otherwise, $Q(a, f) = 0$.
2. At every frame t, a weighted summation $S(t)$ of $Q(t, f)$ is calculated by
$$S(t) = \sum_{f=1}^{2048} F(f)\, Q(t, f),$$
where $F(f)$ is a lowpass filter that is determined as shown in Fig. 2 according to the frequency characteristics of typical bass or snare drums.
3. Each onset time is given by the peak time found by peak-picking in $S(t)$. $S(t)$ is linearly smoothed with a convolution kernel before its peak time is calculated.

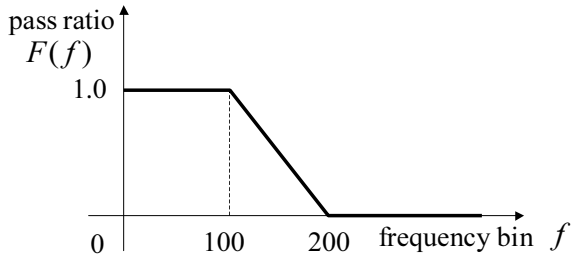

Figure 2: *Function of the lowpass filter according to the frequency characteristics of typical bass and snare drums.*

2.2. Seed Template and Spectrum Excerpt Preparation

The *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, the 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 $T_S(t, f)$ ($1 \leq t \leq 15$ [frames], $1 \leq f \leq 2048$ [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.

On the other hand, a 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. 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 $F(f)$:

$$\begin{aligned} \acute{T}_g(t, f) &= F(f)\, T_g(t, f), \\ \acute{P}_i(t, f) &= F(f)\, P_i(t, f). \end{aligned}$$

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 the 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

$$Corr(j) = \sum_{t=1}^{15} \sum_{f=1}^{2048} \acute{T}_g(t, f)\, \acute{P}_{i,j}(t, f),$$

   where $\acute{P}_{i,j}(t, f) = F(f)\, P_{i,j}(t, f)$.

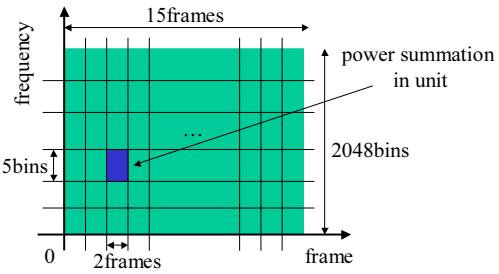

Figure 3: *Quantization at a lower time-frequency resolution for our improved log-spectral distance measure.*

2. The best index J is determined as an index j that maximizes $Corr(j)$.

$$J = \underset{j}{\operatorname{argmax}}\, Corr(j).$$

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

2.3. Excerpt Selection

To select a set of spectrum excerpts P_i that are similar to the template T_g, we propose an *improved log-spectral distance measure*. 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 (the ratio is 0.1 in this paper). 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 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 Fig 3, the time and frequency resolution after the quantization is 2 [frames] (20 [ms]) and 5 [bins] (54 [Hz]), respectively. The D_i between $T_g (T_S)$ and P_i is defined as

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

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

$$\begin{aligned} \hat{T}_g(\hat{t}, \hat{f}) &= \sum_{t=2\hat{t}-1}^{2\hat{t}} \sum_{f=5\hat{f}-4}^{5\hat{f}} \acute{T}_g(t, f), \\ \hat{P}_i(\hat{t}, \hat{f}) &= \sum_{t=2\hat{t}-1}^{2\hat{t}} \sum_{f=5\hat{f}-4}^{5\hat{f}} \acute{P}_i(t, f). \end{aligned}$$

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

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

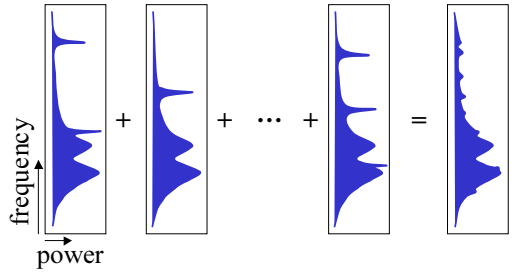

Figure 4: *Updating the template by calculating the median of selected spectrum excerpts.*

2.4. Template Refinement

As is shown in Fig. 4, the median of all the selected spectrum excerpts is calculated and the updated (refined) template T_{g+1} is obtained by

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

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.

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 [2]. 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 Fig. 5. 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

The weight function w is defined as

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

where T_A is the adapted template and $F(f)$ is the low-pass filter function depicted in Fig. 2. The weight function represents the magnitude of spectral characteristic at each frame t and frequency f in the adapted template.

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 template because it cannot judge whether a spectrum excerpt includes the template.

To calculate the loudness difference between a 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 point is calculated. Next, the power difference δ_i at each frame is calculated by using η_i at that frame. 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 loudness difference 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, determined as frequencies where $w(t, f_{t,k})$ is the k-th largest at frame t. The power difference $\eta_i(t, f_{t,k})$ at t and $f_{t,k}$ is calculated as

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

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

$$\begin{aligned} \delta_i(t) &= \min_k \eta_i(t, f_{t,k}), \\ K_i(t) &= \operatorname*{argmin}_k \eta_i(t, f_{t,k}). \end{aligned}$$

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

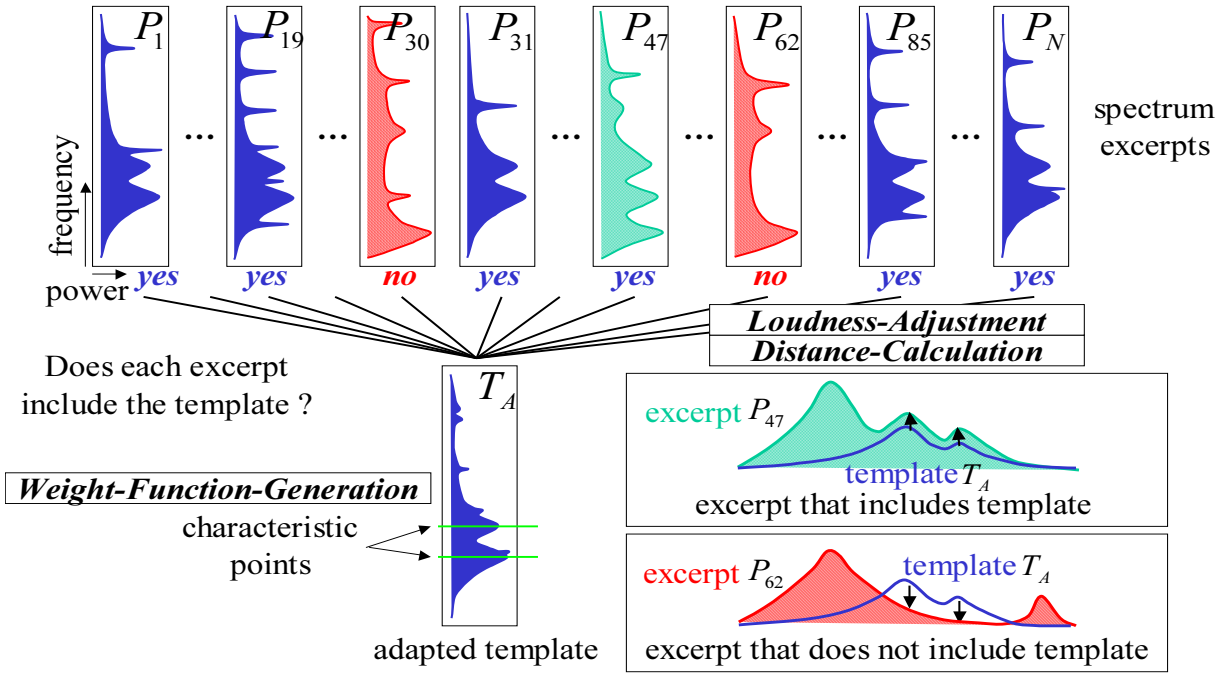

Figure 5: *Overview of template-matching method (matching adapted template with all spectrum excerpts).*

3. The loudness difference Δ_i is calculated as

$$\Delta_i = \frac{\sum_{\{t|\delta_i(t)>\Theta_\delta\}} \delta_i(t)\, w(t, f_{t,K_i(t)})}{\sum_{\{t|\delta_i(t)>\Theta_\delta\}} w(t, f_{t,K_i(t)})}.$$

Let P_i' be an adjusted spectrum excerpt after the loudness adjustment, determined as

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

3.3. Distance Calculation

The distance between the adapted template T_A and an adjusted spectrum excerpt P_i' is calculated by using an extended version of the Goto's distance measure [2]. 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

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

where $\gamma_i(t,f)$ is the local distance between T_A and P_i' at t and f. Ψ is a negative constant to make this 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 weight function w:

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

To determine whether the targeted drum played at P_i', the distance Γ_i is compared with a threshold Θ_Γ. If Γ_i is smaller than Θ_Γ, 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.

4.1. Experimental Conditions

We tested our method on excerpts of five songs included in the popular music database *RWC-MDB-P-2001* developed by Goto *et al.* [3]. Each excerpt was taken from the first minute of a song. The songs we used included sounds of vocals and various instruments in addition to drums as songs in commercial CDs do. Seed templates were created from solo tones included in the musical instrument sound database *RWC-MDB-I-2001* [4]. All data were sampled at 44.1 kHz with 16 bits. The same thresholds were used in the identification of bass drum and snare drums as:

$$\begin{aligned} R_\delta &= 7 \text{ [frames]}, \quad \Psi = \Theta_\delta = -10 \text{ [dB]}, \\ \Theta_\Gamma &= 5000. \end{aligned}$$

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

$$\begin{aligned} \text{recall rate} &= \frac{\text{the number of correctly detected onsets}}{\text{the number of actual onsets}}, \\ \text{precision rate} &= \frac{\text{the number of correctly detected onsets}}{\text{the number of onsets detected by matching}}, \\ \text{F-measure} &= \frac{2 \cdot \text{recall rate} \cdot \text{precision rate}}{\text{recall rate} + \text{precision rate}}. \end{aligned}$$

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

Table 1: *Experimental results for five musical pieces in RWC-MDB-P-2001.*

piece number	method	bass drum			snare drum		
		recall rate	precision rate	F-measure	recall rate	precision rate	F-measure
No.6	base	25.5 % (28/110)	68.3 % (28/41)	0.37	81.0 % (51/63)	83.6 % (51/61)	0.82
	adapt	57.3 % (63/110)	84.0 % (63/75)	0.68	98.4 % (62/63)	100 % (62/62)	0.99
No.11	base	53.8 % (28/52)	100 % (28/28)	0.70	21.6 % (8/37)	66.7 % (8/12)	0.33
	adapt	100 % (52/52)	100 % (52/52)	1.00	94.6 % (35/37)	97.2 % (35/36)	0.96
No.30	base	19.2 % (25/130)	89.3 % (25/28)	0.31	25.7 % (18/70)	90.0 % (18/20)	0.40
	adapt	93.1 % (121/130)	93.8 % (121/129)	0.93	97.1 % (68/70)	100 % (68/68)	0.99
No.50	base	92.4 % (61/66)	93.8 % (61/65)	0.93	91.7 % (99/108)	91.7 % (99/108)	0.92
	adapt	97.0 % (64/66)	87.7 % (64/73)	0.92	61.1 % (66/108)	94.3 % (66/70)	0.74
No.52	base	86.3 % (113/131)	95.8 % (113/118)	0.90	97.4 % (76/78)	93.8 % (76/81)	0.96
	adapt	93.9 % (117/131)	90.4 % (117/128)	0.92	88.5 % (69/78)	97.2 % (69/71)	0.93
average	base	51.1 % (255/489)	91.1 % (255/280)	0.66	70.8 % (252/356)	89.4 % (252/282)	0.79
	adapt	86.5 % (423/489)	91.0 % (423/465)	0.89	88.5 % (300/356)	87.3 % (300/307)	0.90

4.2. Results of Drum Sound Identification

Table 1 shows the 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. The number of adaptive iterations is three. These results showed the effectiveness of the *adapt* method: the template-adaptation method improved the F-measure of identifying bass drum from 0.66 to 0.89 and that of identifying snare drum from 0.79 to 0.90 on average of the five 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 most musical pieces, the recall rate was significantly improved in the *adapt* method. The *base* method detected only a few onsets in some pieces (e.g., No. 11 and No. 30) because the distance between an unadapted seed template and spectrum excerpts was not appropriate. 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.

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 overlaped spectral characteristic points of the bass drum in those excerpts.

5. Conclusion

In this paper, we have described a method that can detect onset times of bass and snare drums in real-world CD recordings containing polyphonic musical audio signals. Even if drum sounds prepared as *seed templates* are different from ones used in a musical piece, our template-adaptation method can adapt the templates to the piece through the iterative adaptation. By using the adapted templates, our template-matching method then detects all the onset times of those drum sounds in the piece by the improved Goto's distance measure. Our experimental results have shown that the adaptation method significantly improved the F-measure of identifying bass and snare drums. In the future, we plan to extend our method to identify other drum sounds and various non-harmonic sounds.

6. 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, the Sound Technology Promotion Foundation, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan).

7. References

[1] Eronen, A. and Klapuri, A., "Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features," *ICASSP*, 753–754, 4, 2000.

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

[3] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R., "RWC Music Database: Popular, Classical, and Jazz Music Databases," *ISMIR*, 287–288, 2002.

[4] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R., "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," *ISMIR*, 229–230, 2003.

[5] Herrera, P., Yeterian, A. and Gouyon, F., "Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques," *ICMAI*, LNAI2445, 4, 49–80, 2002.

[6] Kashino, K., and Murase, H., "A Sound Source Identification System for Ensemble Music Based on Template Adaptation and Music Stream Extraction," *Speech Communication*, 27, 337–349, 1999.

[7] Martin, K. D., "Musical Instrumental Identification: A Pattern-Recognition Approach," *136th meeting of ASA*, 1998.

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