AdaMast: A Drum Sound Recognizer based on Adaptation and Matching of Spectrogram Templates

Kazuyoshi Yoshii† **Masataka Goto**‡ **Hiroshi G. Okuno**†

†Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
‡National Institute of Advanced Industrial Science and Technology (AIST)
Umezono, Tsukuba, 305-8568, Japan

yoshii@kuis.kyoto-u.ac.jp m.goto@aist.go.jp okuno@i.kyoto-u.ac.jp

ABSTRACT

This paper describes a template-matching-based system, called *AdaMast*, that detects onset times of the bass drum, snare drum, and hi-hat cymbals in polyphonic audio signals of popular songs. AdaMast uses the power spectrograms of the drum sounds as templates. However, there are two main problems in transcribing drum sounds in the presence of other sounds. The first problem is that actual drum-sound spectrograms cannot be prepared as templates beforehand for each song. The second problem is that power spectrograms of sound mixtures including the drum sound are greatly different from the template (pure drum-sound spectrogram). To solve the first problem, a template-adaptation algorithm is built into AdaMast. To solve the second problem, a distance measure used in the template matching is designed to be robust to the spectral overlapping of other sounds. The test results in Audio Drum Detection Contest were 72.8%, 70.2%, and 57.4% in transcribing the bass drums, snare drums, and hi-hat cymbals, respectively, and AdaMast won the contest.

Keywords: Drum sound recognition, spectrogram template, template adaptation, template matching.

1 INTRODUCTION

From the viewpoint of the methodology, drum transcription methods are roughly categorized into three types: feature-based classification methods, sound source separation methods, and template-based detection methods. In addition, those methods can be also categorized by focusing on the complexity of input audio signals: solo tones, drum tracks, or musical pieces such as popular songs.

Feature-based classification methods use acoustic feature models trained with database. Herrera et al. (2002) compared conventional classifiers in the experiments of identifying solo drum sounds. To transcribe drum sounds in drums-only audio signals, the use of N-grams (Paulus and Klapuri (2003a)), probabilistic models (Paulus and Klapuri (2003b)), or HMM&SVM (Gillet and Richard (2004)) was proposed. To identify drum sounds extracted from polyphonic audio signals, Van Steelant et al. (2004) reported on the effectiveness of SVM. Sandvold et al. (2004) proposed a feature-model adaptation method that is robust to the distortion of features since the feature distortion caused by other sounds is a main problem.

Sound source separation methods, which are commonly used, originated from the spectrogram decomposition formulation in ISA (Independent Subspace Analysis, Casey and Westner (2000)). To transcribe drum sounds in audio signals of drum tracks, various assumptions are made in decomposing a single music spectrogram into multiple spectrograms of drum instruments; ISA (FitzGerald et al. (2002); Uhle et al. (2003)) assumes the statistical independence of sources, NMF (Non-negative Matrix Factorization, Paulus and Klapuri (2005)) assumes their non-negativity, and sparse coding (Virtanen (2003)) assumes their non-negativity and sparseness. Further developments were made by FitzGerald et al. (2003b,a). They proposed PSA (Prior Subspace Analysis, FitzGerald et al. (2003b)), which assumes the prior frequency characteristics of drum sounds, and applied it to transcribe drum sounds in the presence of harmonic sounds (FitzGerald et al. (2003a)). For the same purpose, Dittmar and Uhle (Dittmar and Uhle (2004)) adopted non-negative ICA that considers the non-negativity of sources. To attain good separation results, it is necessary to estimate the number of sources, but it is difficult to precisely estimate it in general. In addition, it is also necessary to identify the drum-instrument label of each separated signal or spectrogram and detect the onset times of the target drum sound.

Template-based detection methods are based on a typical pattern recognition approach — the distance between a template and an input pattern is calculated. Goto and Muraoka (1994) proposed a template-matching method that used spectrogram templates, and transcribe drum sounds in drum-track audio signals consisting of drums only. Gouyon et al. (2000) proposed a method that classifies mixed sounds extracted from polyphonic audio signals into two categories (bass and snare drums). To detect those drum sounds to be classified, they proposed a template-adaptation method using waveform templates. It can deal with drum-sound variations found in musical pieces. Zils et al. (2002) extended Gouyon's method, and tried the extraction of bass and snare drum sounds from CD recordings. In general, it is difficult to deal with the difference between a template and an actual pattern used in a musical piece. To deal with this difference in the time-frequency space and achieve more robust performance, AdaMast (Yoshii et al. (2004, 2005)) was developed by integrating Goto's matching method and Zils' adaptation method.

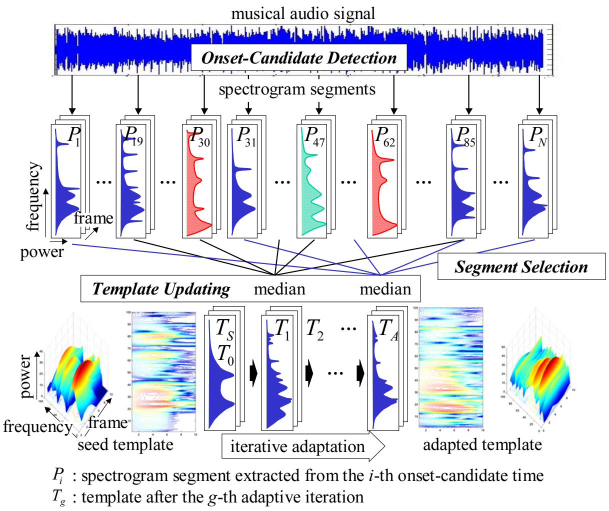

Figure 1: Overview of template-adaptation method.

2 ARCHITECTURE

AdaMast is a template-based drum sound recognition system. We briefly describe a recognition algorithms of it, and explain our system-design policy.

2.1 Algorithms

AdaMast is composed of successive template-adaptation and template-matching parts. They use the power spectrograms of the drums as templates. Detailed explanation of the core algorithms is described in our previous paper (Yoshii et al. (2004)). Further extensions (e.g., automatic thresholding, harmonic-structure suppression) are described in another paper (Yoshii et al. (2005)).

2.1.1 Template Adaptation

The purpose of the adaptation method is to construct a spectrogram template that is adapted to its corresponding drum-sound spectrogram in the polyphonic audio signal of a target musical piece. Before starting the adaptation, we prepare power-spectrogram templates (we call *seed templates*) for the bass drum, snare drum, and hi-hat cymbals, respectively; three templates in total. To analyze audio signals sampled at 44.1 [kHz], we used short-time Fourier transform (STFT) with a Hanning window (4096 points) with a shifting interval of 441 points. The time-length of the templates is set to 10 [frames]. To adapt the seed templates to the actual drum-sound spectrograms, we extended Zils' adaptation method (Zils et al. (2002)) to the time-frequency domain.

Our method is based on an iterative adaptation algorithm. An overview is shown in Figure 1. First, *onset-candidate detection* stage roughly detects onset candidates in the input audio signal of a musical piece. Starting from each onset candidate, a spectrogram segment with a fixed time length is extracted from the power spectrogram of the input audio signal. Then, using the seed template and all the spectrogram segments, the iterative algorithm successively applies *segment selection* and *template updating* to obtain an adapted template.

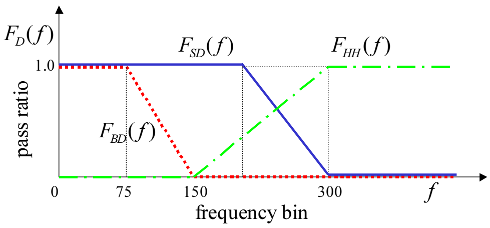

Figure 2: Filter functions F_{BD}, F_{SD}, and F_{HH} represent typical frequency characteristics of bass drums, snare drums, and hi-hat cymbals.

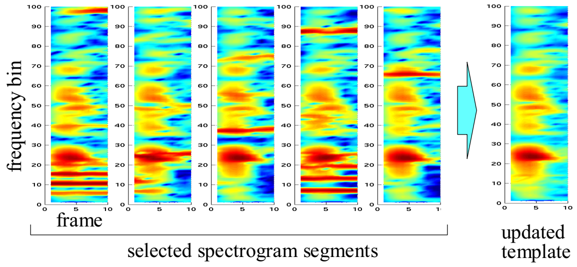

Figure 3: Template updating by collecting median power at each frame and each frequency bin for selected spectrogram segments.

Let T_i $(i = 1 \cdots N_D)$ denote a frame detected as an onset candidate and P_i denote a spectrogram segment extracted from T_i (N_D is the number of detected onset candidates). These selection and updating work as follows:

1. *Segment selection* calculates the reliability R_i that spectrogram segment P_i includes the drum sound spectrogram. The reliability is defined as the reciprocal of the Euclidean spectral distance:

$$R_i = \frac{1}{\sqrt{\sum_{t=1}^{10} \sum_{f=1}^{2048} \left(\acute{T}_g(t,f) - \acute{P}_i(t,f)\right)^2}}, \quad (1)$$

   where T_g is the template after the g-th adaptive iteration. In practice, we used a modified version of this measure. $\acute{T}_g$ and $\acute{P}_i$ are low-pass/high-pass filtered spectrograms:

$$\acute{T}_g(t,f) = F_D(f)\, T_g(t,f), \quad (2)$$
$$\acute{P}_i(t,f) = F_D(f)\, P_i(t,f), \quad (3)$$

   where $F_D(f)$ $(\{D|BD, SD, HH\})$ is a low-pass or high-pass filter function, as shown in Figure 2. We assume that it represents the typical frequency characteristics of bass drum sounds (BD), snare drum sounds (SD), and hi-hat cymbal sounds (HH). Spectrogram segments with high reliabilities are then selected; this selection is based on a fixed ratio to the total number of segments.

2. *Template updating* then reconstructs an updated template by estimating the power that is defined, at each time and each frequency bin, as the median power

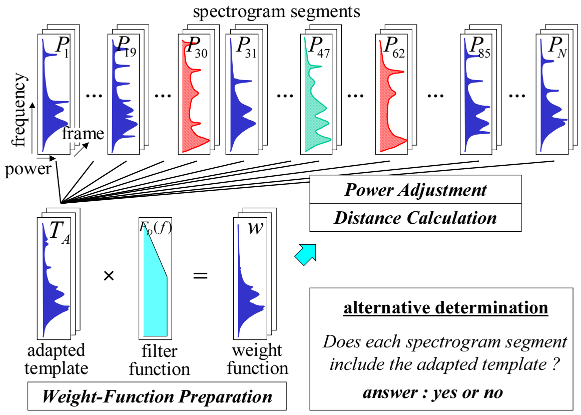

Figure 4: Overview of template-matching method.

among the selected spectrogram segments (Figure 3). The median operation can suppress harmonic components in the updated template. The template is thus adapted to the current piece and used for the next adaptive iteration. The updated template, $\acute{T}_{g+1}$, is weighted by filter function F_D and is obtained by

$$\acute{T}_{g+1}(t,f) = \operatorname*{median}_{1 \leq i \leq N_S} \acute{P}^{(i)}(t,f), \tag{4}$$

where $P^{(i)}$ $(i = 1, \cdots, N_S)$ are the spectrogram segments selected by *segment selection*. N_S is the number of selected spectrogram segments, which is empirically set to $0.1 \times N_D$ in transcribing any drum.

2.1.2 Template Matching

This method detects all the onset times of the drum sounds in the polyphonic audio signal, even if other musical instrument sounds overlap the drum sounds. To find the actual onset times, this method determines *whether the drum sound actually occurs at each onset candidate time*, as shown in Figure 4. The matching distance is calculated using Goto's distance measure (Goto and Muraoka (1994)). Since this method focuses on whether the adapted template is included in a spectrogram segment, it can calculate an appropriate distance even if the drum sound is overlapped by other musical instrument sounds.

1. *Weight-function preparation* generates a function that represents the spectral saliency of each frequency component in the adapted template. This function is used for selecting characteristic frequency bins. The weight function w is defined as

$$w(t,f) = F_D(f)\, T_A(t,f), \tag{5}$$

where T_A is the adapted template and F_D is the filter function.

2. *Power adjustment* calculates the power difference between the template and each spectrogram segment by focusing on the characteristic frequency bins. If the power difference is larger than a threshold, it judges that the drum sound spectrogram does not appear in that segment and does not execute the subsequent processing. Otherwise, the power of that segment is adjusted to compensate for the power difference. Let P_i' be a power-adjusted spectrogram segment.

3. *Distance calculation* calculates the spectral distance between adapted template T_A and each P_i'. If $P_i'(t,f)$ is larger than $T_A(t,f)$, Goto's distance measure regards $P_i'(t,f)$ as a mixture of frequency components not only of the drum sounds but also of other musical instrument sounds. In other words, if we determine that $P_i'(t,f)$ includes $T_A(t,f)$, then the local distance at frame t and frequency bin f is minimized. Therefore, the local distance is defined as

$$\gamma_i(t,f) = \begin{cases} 0 & \text{if } \left(P_i'(t,f) - T_A(t,f) \geq \Psi\right), \\ 1 & \text{otherwise,} \end{cases} \tag{6}$$

where Ψ is a negative constant, which is set to -12.5 [dB] in this paper.

Total distance Γ_i is calculated by integrating local distance γ_i in the time-frequency domain, weighted by w:

$$\Gamma_i = \sum_{t=1}^{10} \sum_{f=1}^{2048} w(t,f)\, \gamma_i(t,f). \tag{7}$$

To determine whether the target drum sound occurred at a time corresponding to spectrum segment P_i', distance Γ_i is compared with threshold Θ_Γ. If $\Gamma_i < \Theta_\Gamma$, we conclude that the target drum sound occurred. The Θ_Γ is automatically determined using Otsu's thresholding algorithm Otsu (1979).

2.2 Design Policy

We use two different distance measures between the template adaptation and matching methods. In the adaptation method, it is desirable to detect only semi-pure drum sounds that have little overlap with other sounds. Those drum sounds tend to result in a good adapted template that includes few frequency components of other sounds. Because it is not necessary to detect all the onset times of the target drum sounds, the distance measure does not need to consider spectral overlapping of other sounds. In the matching method, on the other hand, we used Goto's distance measure because it is necessary to exhaustively detect all the onset times even if the target drum sounds are overlapped by other sounds.

3 EVALUATION

To fairly evaluate multiple drum-sound transcription algorithms and compare their results, Audio Drum Detection Contest was held as a track of MIREX2005.

3.1 Conditions

The drum types to be transcribed are the bass drum (BD), snare drum (SD), and hi-hat cymbals (HH). The test set was consisted of 30-[s] fragments and entire musical performances which were sampled from many genres. A representative random subset (20% of all available files) of the data was available to all participants in advance of the evaluation. Participants should only use the data made available to all participants by the organizers. To evaluate results, F-measure (harmonic mean of the recall and precision rates) was calculated for each of three drum types, resulting in three F-measure scores and their average score.

Table 1: Evaluation results (F-measures and runtime).

Participant	Total	BD	SD	HH	Runtime*
Yoshii, K.	0.670	0.728	0.702	0.574	8534 [s]
Tanghe, K.	0.611	0.688	0.555	0.601	1337 [s]
Dittmar. C.	0.588	0.606	0.581	0.585	673 [s]
Paulus, J.	0.499	0.527	0.430	0.587	1137 [s]
Gillet, O.	0.443	0.598	0.428	0.334	21248 [s]

*We cannot directly compare the runtime because different machines with different OSs and CPUs were used.

Table 2: Feature-based classification methods.

Participant	Feature sets	Compression	Decision
Tanghe, K.	MFCC etc.	-	SVM
Dittmar. C.	Band energy etc.	LDA	*k*NN
Paulus, J.	MFCC etc.	PCA	HMM
Gillet, O.	MFCC etc.	-	SVM

3.2 Observation

The test results are shown in Table 1. AdaMast yielded the best total score and became a winner of this-year's contest. Although the runtime of AdaMast is comparatively long, AdaMast is fast enough to complete the processing within the playing time of the target signal. It is interesting that only AdaMast is based on a template-matching method, which uses drum-sound spectrograms as templates.

On the contrast, the four systems other than AdaMast use feature-based classification methods. Table 2 shows an overview of these systems. The features they used are similar to those proposed by Herrera et al. (2002). In general, LDA is more suitable than PCA to compress the feature dimension for the purpose of improving the classification capability. Since Tanghe's and Gillet's systems use SVM, the compression is not necessary. To extract quasi (semi-pure) drum-sound spectrograms, Dittmar's system used non-negative ICA and a method motivated by our template-updating concept, and Gillet's system used noise-space-projection method. Among these four systems, Tanghe's one yielded the best score. Although the high generalization capability of SVM can be effectively utilized by extracting many kinds of features, there still remains the generalization limitation.

In our observation, we think the existence of an adaptation mechanism is more critical than the methodological difference. If the template-adaptation part of AdaMast is disabled, the results will become much worse. The top three systems in Table 2 do not have adaptation mechanisms of general feature models. Sandvold et al. (2004) reported on the large improvement of the performance by using localized feature models. However, to create the reliable localized models, they used correct drum-type labels. In other words, a criterion for evaluating the reliability of each onset from a different viewpoint is needed. A solution of this difficult problem is included in future work. In addition, it is necessary to address musical pieces which do not include the sounds of all the drum types.

ACKNOWLEDGEMENTS

Many thanks to Koen Tanghe for proposing and coordinating Audio Drum Detection Contest. We appreciate the continuous effort of MIREX2005 committee.

References

M.A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proc. Int. Computer Music Conference (ICMC)*, 2000.

C. Dittmar and C. Uhle. Further steps towards drum transcription of polyphonic music. In *Proc. Audio Engineering Society (AES), 116th Convention*, 2004.

D. FitzGerald, E. Coyle, and B. Lawlor. Sub-band independent subspace analysis for drum transcription. In *Proc. Int. Conf. Digital Audio Effects (DAFX)*, pages 65–69, 2002.

D. FitzGerald, B. Lawlor, and E. Coyle. Drum transcription in the presence of pitched instruments using prior subspace analysis. In *Proc. Irish Signals and Systems Conference (ISSC)*, pages 202–206, 2003a.

D. FitzGerald, B. Lawlor, and E. Coyle. Prior subspace analysis for drum transcription. In *Proc. Audio Engineering Society (AES), 114th Convention*, 2003b.

O. Gillet and G. Richard. Automatic transcription of drum loops. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 269–272, 2004.

M. Goto and Y. Muraoka. A sound source separation system for percussion instruments. *IEICE Trans. D-II*, J77-D-II(5): 901–911, May 1994.

F. Gouyon, F. Pachet, and O. Delerue. On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proc. COST-G6 Conference on DAFX*, 2000.

P. Herrera, A. Yeterian, and F. Gouyon. Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques. In *Proc. ICMAI, LNAI2445*, pages 69–80, 2002.

N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Sys., Man, and Cybern.*, 6(1):62–66, 1979.

J. Paulus and A. Klapuri. Conventional and periodic N-grams in the transcription of drum sequences. In *Proc. Int. Conf. Multimedia and Expo (ICME)*, pages 737–740, 2003a.

J. Paulus and A. Klapuri. Model-based event labeling in the transcription of percussive audio signals. In *Proc. Int. Conf. Digital Audio Effects (DAFX)*, pages 73–77, 2003b.

J. Paulus and A. Klapuri. Drum transcription with non-negative spectrogram factorisation. In *Proc. European Signal Processing Conference (EUSIPCO) (in press)*, 2005.

V. Sandvold, F. Gouyon, and P. Herrera. Percussion classification in polyphonic audio recordings using localized sound models. In *Proc. ISMIR*, pages 537–540, 2004.

C. Uhle, C. Dittmar, and T. Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, pages 843–848, 2003.

D. Van Steelant, K. Tanghe, S. Degroeve, B. De Baets, M. Leman, and J.-P. Martens. Classification of percussive sounds using support vector machines. In *Proc. Benelearn*, 2004.

T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *Proc. Int. Computer Music Conference (ICMC)*, pages 231–234, 2003.

K. Yoshii, M. Goto, and H.G. Okuno. Automatic drum sound description for real-world music using template adaptation and matching methods. In *Proc. ISMIR*, pages 184–191, 2004.

K. Yoshii, M. Goto, and H.G. Okuno. Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates after suppressing harmonic structure. *IEEE Trans. on Speech and Audio Processing (in submitted)*, 2005.

A. Zils, F. Pachet, O. Delerue, and F. Gouyon. Automatic extraction of drum tracks from polyphonic music signals. In *Proc. Int. Conf. Web Delivering of Music (WEDELMUSIC)*, pages 179–183, 2002.