A Novel Approach to Separation of Musical Signal Sources by NMF

Sakurako Yazawa
Graduate School of Systems and Information Engineering, University of Tsukuba, Japan

Masatoshi Hamanaka
Department of Clinical System Onco-Informatics, Kyoto University, Japan

Takehito Utsuro
Faculty of Engineering, Information and Systems, University of Tsukuba, Japan

***Abstract*—This paper proposes a method to separate polyphonic music signal into signals of each musical instrument by NMF: Non-negative Matrix Factorization based on preservation of spectrum envelope. Sound source separation is taken as a fundamental issue in music signal processing and NMF is becoming common to solve it because of its versatility and compatibility with music signal processing. Our method bases on a common feature of harmonic signal: spectrum envelopes of musical signal in close pitches played by the harmonic music instrument would be similar. We estimate power spectrums of each instrument by NMF with restriction to synchronize spectrum envelope of bases which are allocated to all possible center frequencies of each instrument. This manipulation means separation of components which refers to tones of each instrument and realizes both of separation without pre-training and separation of signal including harmonic and non-harmonic sound. We had an experiment to decompose mixture sound signal of MIDI instruments into each instrument and evaluated the result by SNR of single MIDI instrument sound signals and separated signals. As a result, SNR of lead guitar and drums approximately marked 3.6 and 6.0 dB and showed significance of our method.**

Keywords—NMF, spectrum envelope preservation, unsupervised learning, polyphonic music signal separation,

I. Introduction

The goal of our work is to extract signals of a specific musical instrument from musical piece signal data. If it becomes possible to highly precisely extract sound sources of any specific musical instrument, then one is able to extract and operate musical parts and musical piece. It is further expected that it enables certain operation of each musical part's volume, which then makes much higher level operation such as arrangement of musical piece much easier than before.

We focus on a sound source of rock band formation and propose a sound source separation technique for a sound source of rock band formation. In this paper, we assume that the sound source of rock band formation is that of four part formation, which consists of the lead guitar, the backing guitar, the bass, and the drums. With this sound source, the sound signal is a music sound signal of performance sound which is a mixture of harmonic instruments and non-harmonic instruments. This paper proposes how to separate musical sound signal of this type into the sound signal of each musical instrument.

The performance of the rock band formation is popular among music fans of young generation, who often like to imitate the performance of certain famous rock bands. Through imitating the performance of certain famous rock bands and the nuance of their performance, it is necessary for those beginners, not only to simply follow note information, but to learn how to vary strength in musical sound, to add certain expression, and to change melodies from part to part by changing performance. However, it is usually difficult for beginners to distinguish the sound of a specific part from the sound source with more than one electric guitars for example. Thus, those musical beginners definitely have strong needs for a technique of separating a sound source of specific parts.

Recently, in the field of the sound source separation, Non-negative Matrix Factorization (NMF) has been intensively studied. NMF is a technique that interprets the spectrogram of the sound signal as a non-negative matrix, and approximate it as the product of a base vector group and a temporal activation group. NMF is often employed in certain formalization of music sound signals for the following two reasons: (1) it can express musical instruments through certain characteristics, and (2) additivity of the power spectrum of musical instruments approximately holds. So far, there exist several previous work on the sound source separation using NMF including the followings: an approach of utilizing the musical score information as the supervision signal [1], that of modeling harmonic and non-harmonic structures and the reverberation information, aiming at sound source separation and replacement of an instrument performance phrase [2], that of modeling time-varying spectral patterns of musical instruments through NMF framework [3].

However, those previous work are limited in that although some of them satisfy just one of the following two requirements, any single technique by itself does not satisfy both of the following two requirements: namely, (i) unsupervised method of sound source separation without supervision information, (ii) realizing separation of harmonic and non-harmonic instruments. As for the first requirement, supposing the case where one wants to separate the sound source of a musical piece that are unknown to him/her, it is common that he/she is not given information such as musical scores in advance. As for the second requirement, the fundamental technique of the sound source separation by NMF identifies temporal activation of a base vector group, while it lacks information on which musical instrument each base actually expresses.

In order to overcome those two obstacles, this paper proposed a novel sound separation method based on NMF. Our method bases on a common feature of harmonic signal: spectrum envelopes of musical signal in close pitches played by the harmonic music instrument would be similar. We estimate

978-1-4799-2186-7/14/$31.00 ©2014 IEEE

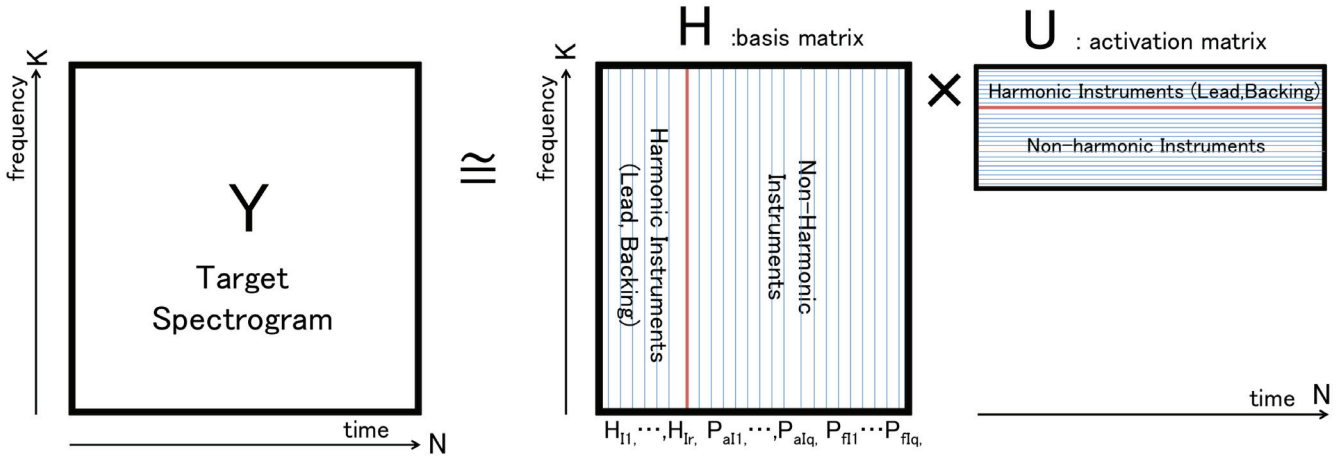

Fig. 1. Proposed Model of Separating Musical Signal Sources by NMF

power spectrums of each instrument by NMF with restriction to synchronize spectrum envelope of bases which are allocated to all possible center frequencies of each instrument. This manipulation means separation of components which refers to tones of each instrument and realizes both of separation without pre-training and separation of signal including harmonic and non-harmonic sound. We have an experiment to decompose mixture sound signal of MIDI instruments into each instrument and evaluate the result by SNR of single MIDI instrument sound signals and separate signals. As a result, SNR of lead guitar and drums approximately mark 3.6 and 6.0 dB and show significance of our method.

II. Related Work

Some previous work on sound source separation for mixed sound source of more than one instruments models the acoustic characteristics of each musical instrument and takes a parametric approach of updating and optimizing certain parameters through EM algorithm. Among them, Goto [4] proposed PreFEst which estimates a melody line and a bass line from the mixed sound source of more than one musical instruments. Itoyama [5] defined a model of harmonics and non-harmonics structure, and by utilizing the MIDI sound source as a supervision information, proposed a technique of iteratively converging the model towards the target signal through EM algorithm.

Among those work on sound source separation using NMF, Uchiyama [6] proposed how to separate the sound source of harmonic and non-harmonic instruments based on the fact that, in spectrum, harmonic instruments tend to be smooth in temporal direction and to change drastically in frequency direction, while non-harmonic instruments tend to be smooth in frequency direction and to change drastically in temporal direction. This fact is also advantageous in our work for the purpose of separating harmonic and non-harmonic mixture of sound signal. However, this approach has limitation when the mixture of sound signal contains more than one harmonic instruments. Furthermore, even harmonic musical instruments contain non-harmonic sound such as attack sounds of the guitar, while even non-harmonic musical instruments contain harmonic sound such as the interval of the tom. Thus, it is not enough to simply separate harmonic and non-harmonic mixture of sound signal, but it is necessary to invent a model which can express both harmonic and non-harmonic sound of every such musical instrument.

Another approach [7] utilized a base for each pitch name and initialized them as a comb shape in the frequency space so that they have harmonic structure, where they proposed how to detect notes with chromatic scales. Although this approach is also effective in our task of modeling separation of harmonic instruments, it again has limitation when the mixture of sound signal contains more than one harmonic instruments.

Schmidt [8] studied how to separate sound source of more than one harmonic musical instruments under the following two assumptions: (a) harmonic structures of instruments' sound are shift immutable in logarithmic frequency, (b) the power ratio of the harmonic overtone is fixed even if the central frequency changes. This approach considers tones of harmonic musical instruments as characteristics when separating the sound sources, while it is limited in that the tone of one musical instrument is fixed for all the pitches. This assumption is not appropriate in our work since musical instruments having several octaves range have quite different tones according to high and low pitches. Thus, it is required in our work to model continuously changing tones according to the height of pitches in the framework of music signal source separation.

III. The Proposed Approach

This sections roughly describes our approach to music signal source separation for the mixture of harmonic and non-harmonic musical instruments with the performance in rock band formation. The musical instruments used in the performance in rock band formation is mostly fixed to some extent, and is in general with the guitar, the bass, the drum, and the keyboard. However, the tone of each musical instrument greatly varies according to musical pieces. In the case of the guitar, for example, their exist roughly two types of sound, one is clean sound without any sound effect units, while the other is those with certain distortion through sound effect units. Its

sound further changes if it is through a certain equalizer. This situation of the guitar is quite different from those studied in the music source separation of other musical instruments such as piano, the wind instruments, and the violin. In the case of the guitar, it is difficult to provide the model with the acoustic characteristics in advance, which makes it difficult to apply the supervised learning approach. In order to solve this problem, in our proposed framework, we assume that the numbers and the types of each musical instrument are given beforehand, and then formalize bases of NMF under this restriction in order to learn the tone of each musical instrument and then to realize music source separation.

Figure 1 shows the structure of bases. The number of bases varies according to each musical instrument and we specify a certain number for each musical instrument. For each non-harmonic musical instrument, we prepare two bases, one of which is for non-harmonic sound, while the other is for harmonic sound accompanying the non-harmonic sound. For each harmonic musical instrument, we prepare as many bases as the frequency bins included in all the frequency bands that the harmonic musical instrument has. Here, note that it is not sufficient if one just prepares exactly the same number of bases corresponding to the chromatic scales but no more bases. This is because it is possible to play harmonic musical instruments such as the guitar with a high level playing technique such as *bending* (after twanging a string, pulling and pushing the string with fingers and changing the pitch smoothly), which results in continuous changes of the pitch.

IV. The Model of Separating Musical Sound Sources

This section describes the details of the model of separating musical sound sources. In NMF, the iteration continues while minimizing divergence between the estimated matrix and the target spectrogram, where the overview of the model is as shown in Figure 1. First, let Y be a matrix of acoustic signals having those of musical instruments to be separated. Then, we decompose Y into the matrix H with bases for musical instrument parts, and that U with temporal activation for those bases. The framework of sound source separation by NMF is formalized as minimizing the objective function in (1) [9],

$$C(\theta) = D(\theta) + R_p(\theta) + R_h(\theta) + R_b(\theta) + R_l(\theta) \quad (1)$$

where each formula in the right hand side is as listed below:

- $D(\theta)$ divergence between the estimated matrix and the target spectrogram.
- $R_p(\theta)$ the cost function for separating non-harmonic musical instruments.
- $R_h(\theta)$ the cost function that is common for separating harmonic musical instruments.
- $R_b(\theta)$ the cost function for separating backing harmonic musical instruments.
- $R_l(\theta)$ the cost function for separating lead harmonic musical instruments.

A. Separating Non-harmonic Instruments

For each non-harmonic musical instrument, we prepare two bases, one of which is for non-harmonic sound that is smooth in frequency direction, while the other is for harmonic sound accompanying the non-harmonic sound. We prepare the base for the harmonic sound because even non-harmonic musical instruments contain harmonic sound such as the interval of the tom and the snare.

More specifically, we prepare bases with the constraint of having smooth envelope in frequency direction, and other bases with restricted activation. Let I_{pi} $(i = 1, \ldots, q)$ (corresponding to the H matrix P_{aI_i} and P_{fI_i} in Figure 1) denote non-harmonic musical instruments, and P_{ahi} and P_{fhi} denote indices of bases for those non-harmonic musical instruments. Then, the following equations give how to initialize bases of those non-harmonic musical instruments as well as the cost function $R_p(\theta)$ is defined: [10]

$$\begin{aligned} h_{km} &= 1, (m = P_{ahi}, P_{fhi}) \\ R_p(\theta) &= \gamma_{pa} \sum_i \sum_{n=2}^{K} (u_{nP_{ahi}} - u_{n-1P_{fhi}})^2 \\ &\quad + \gamma_{pf} \sum_i \sum_{k=2}^{K} (h_{kP_{fhi}} - h_{k-1P_{fhi}})^2 \end{aligned}$$

Here, the coefficients γ_{pa} and γ_{pf} represent relative weights of the first and the second terms of the cost function. The more smooth the activation is, the smaller the cost function above is designed to be.

B. Separating Harmonic Instruments

The general formalization of separating harmonic instruments overall follows that by Raczyński [7]. In this paper, we classify harmonic musical instruments into those for lead such as with playing main melody and those for backing such as with playing chord. The underlying idea of separating lead and backing is to separate them by changing constraints in NMF, since the lead instrument plays the single tone while the backing instrument plays the chord sound.

Let I_{hi} $(i = 1, \ldots, r)$ (corresponding to the H matrix H_{I_i} in Figure 1) denote harmonic musical instruments, and I_{hi} denote the set of indices of bases for those harmonic musical instruments. Then, the following equation gives how to initialize bases of those harmonic musical instruments.

$$h_{km} = \begin{cases} 1 & (mod(k - \omega_{lowhi}, \xi) = 0,\ \xi \in H_{hi} \\ 0 & (otherwise) \end{cases}$$

Supposing that we are given any possible central frequency that the harmonic musical instrument I_{hi} can play, then, with this initialization, we can prepare bases with comb shape which have all the overtones of the given central frequency. Next, the cost function $R_h(\theta)$ that is common for separating harmonic musical instruments is defined as below:

$$R_h(\theta) = \gamma_h \sum_{i,j} \sum_{m \in H_{hi}} (h_{\phi(m,j)m} - h_{\phi(m-1,j)m-1)})^2$$

where the coefficient γ_h represents a relative weight of the cost function, and $\phi(m, j)$ represents j-th index which satisfies $h_{km} \neq 0$. This cost function is meant to be that the strengths of the overtones of the same level of adjacent bases are designed to be closer in a harmonic musical instrument. This design of the cost function enables the modeling of the continuous changes of the pitch with harmonic musical instruments such as the guitar mentioned in section III.

Next, the following sections describe how to separate backing and lead harmonic instruments.

1) Separating Backing Instruments: Since the backing instrument plays the chord sound, it is probable that it has more than one central frequencies. Here, we assume that those central frequencies are mostly observed within a rather narrow range in the frequency space. By expressing this assumption through a probability distribution, those central frequencies are mostly observed very closely around the frequency bin which gives the expectation on the frequency axis. Considering the observation above, we design the requirement for separating backing instruments as having those activations which minimize their variance. Let H_b ($\subset H_{hi}$) be the set of indices of bases for those backing harmonic musical instruments, then the cost function $R_b(\theta)$ for separating backing harmonic musical instruments is defined as below:

$$R_b(\theta) = \gamma_b \sum_n \Big(\frac{\sum_{m\in H_b} m^2 u_{mn}}{\sum_{m\in H_b} u_{mn}} - \mu^2 \Big)$$

$$\mu = \frac{\sum_{m\in H_b} m u_{mn}}{\sum_{m\in H_b} u_{mn}}$$

where the coefficient γ_b represents a relative weight of the cost function.

2) Separating Lead Instruments: Since the lead instrument plays the main melody, it is probable that it has one specific central frequency. Considering this observation, we design the requirement for separating lead instruments as having those activations which maximize their kurtosis. Let H_l ($\subset H_{hi}$) be the set of indices of bases for those lead harmonic musical instruments, then the cost function $R_l(\theta)$ for separating lead harmonic musical instruments is defined as below:

$$\begin{aligned} R_l(\theta) = -\gamma_l \sum_n \Big\{ \Big(& \sum_{m\in H_l} m^4 u_{mn} - 4 \sum_{m\in H_l} m^3 u_{mn} \\ & + 6\mu^2 \sum_{m\in H_l} m^2 u_{mn} \Big) \Big(\sum_{m\in H_l} u_{mn} \Big)^{-1} \\ & - 3\mu^4 \Big\} \end{aligned}$$

where the coefficient γ_l represents a relative weight of the cost function.

V. Evaluation

This section shows results of two experimental evaluation: the one with the MIDI sound source, and the other with the CD sound source. In the evaluation with the MIDI sound source, we apply the proposed technique of sound source separation to the mixture of the MIDI sound source of each musical instrument, and then evaluate how well the proposed method of sound source separation performs by referring to the MIDI sound source of each musical instrument.

TABLE I. Parameters and their Values in Evaluation

parameter	value
Sampling rate	44,100Hz
STFT window	Hamming (8,192pts)
overlap	7,680
central frequency range (Lead Guitar)	C4(261.43Hz) $\sim$ C7(2,093Hz)
central frequency range (Back Guitar)	C3(130.81Hz) $\sim$ C5(523.25Hz)
central frequency range (Bass)	C2(65.406Hz) $\sim$ C4(261.43Hz)
γ_h	16,384
γ_{pa}	1
γ_{pf}	13,1072
γ_l	1.0×10^{-6}
γ_b	2.0×10^{-3}

TABLE II. Evaluation Results with the MIDI Sound Source [dB]

	LeadG	BackG	Ba	Dr	Ave
t001	5.442	11.495	-0.662	8.843	6.28
t002	8.973	8.154	-0.595	6.464	5.749
t003	2.921	-2.252	-7.733	-5.217	-3.07
t004	7.908	1.145	1.207	6.695	4.239
t005	2.025	-2.091	4.478	7.075	2.872
t006	6.968	4.097	0.485	6.538	4.522
t007	4.967	-31.206	1.243	10.84	-3.539
t008	-2.956	-0.113	7.732	11.971	4.158
t009	0.913	-2.107	1.602	2.156	0.641
t010	-1.61	-0.537	2.56	4.258	1.168
Ave	3.555	-1.341	1.032	5.962	

A. Evaluation with the MIDI Sound Source

1) The Procedure: As the MIDI sound source, we use ten pieces of sound signals each of which consists of the MIDI sound sources of the lead guitar, the backing guitar, the bass, and the drums of 10 seconds. After we apply the proposed technique of sound source separation to the mixture of the MIDI sound sources of musical instruments, and have the results of separated sound signals, we evaluate the results with the following SNR:

$$SNR = 10 \log_{10} \frac{Var(O_t)}{Var(S_t)}$$

where O_t denotes the original MIDI sound source for each musical instrument, while S_t denotes the acoustic signal obtained by decoding the difference of the spectrogram of O_t and that obtained from the result of separated sound signals. Here, Table I lists the values of the parameters used in the evaluation of this paper.

2) Evaluation Results: Table II lists the results of evaluation for each of the musical instruments. As can be seen from this result, we achieve especially high performance in separating the lead guitar and the drums. This is partially because we assign a larger number of bases to the lead guitar than to the backing guitar and the bass, which enables more detailed representation of the sound source of the lead guitar. As for the drums, we assign three bases considering the high-hat, the snare, and the bass drum, while it is not easy to decide that we successfully separate those three sound sources by the proposed technique. Thus, it is further necessary for us to invent another technique to separate those sound sources of the non-harmonic musical instruments.

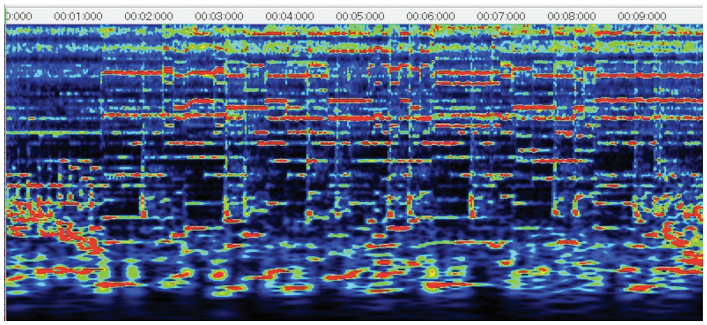

Fig. 2. Spectrogram of the CD Sound Source of Test Data

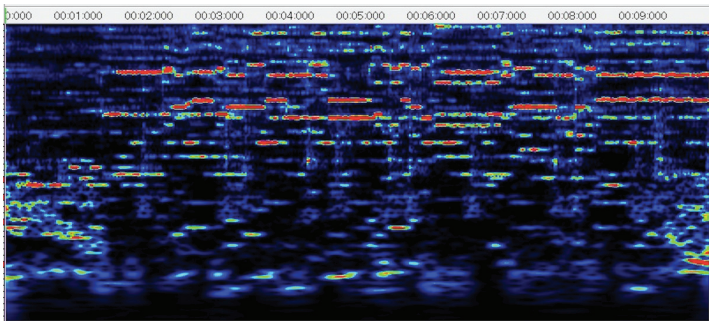

Fig. 3. Spectrogram of the Separated Lead Guitar

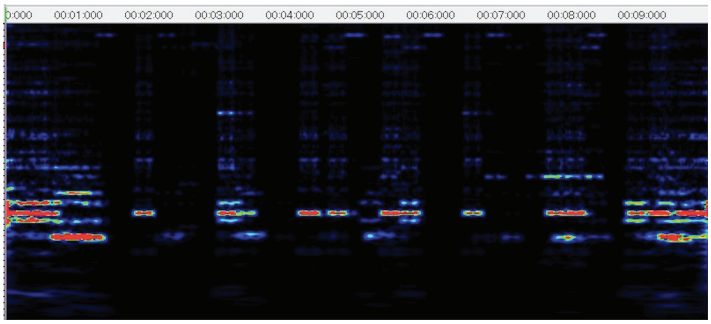

Fig. 4. Spectrogram of the Separated Backing Guitar

B. Evaluation with the CD Sound Source

We applied the proposed method to a CD sound source. The musical piece of this CD sound source is a part with the guitar solo for approximately ten seconds taken from 52 seconds in ”thunder force 666” (composed by Azuma Ruriko) [11]. ”Thunder force 666” is a sound source with a musical piece of hard melodic metal including an electric guitar and drums, which matches the proposed technique. Figure 2 shows the spectrogram of this CD sound source. Figures 3 $\sim$ 6 show the spectrograms of the separated lead guitar, the backing guitar, the bass, and the drums, while Figure 7 shows that of their mixed down data. It is obvious from this result that separation is performed with a precision as high as that of the evaluation with the MIDI sound source.

VI. Conclusion and Future Work

This paper proposed a method to separate polyphonic music signal into signals of each musical instrument by NMF based on preservation of spectrum envelope. Based on the evaluation results presented in this paper, it is definitely required that we intensively evaluate the proposed method with a much larger number of sound sources including those listed in the Metal Songs Top 100.

References

[1] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. Okuno, “Simultaneous realization of score-informed sound source separation of polyphonic musical signals and constrained parameter estimation for integrated model of harmonic and inharmonic structure,” *Journal of Information Processing Society of Japan*, vol. 49, no. 3, pp. 1465–1479, 2008, (in Japanese).

[2] N. Yasuraoka, T. Yoshioka, K. Itoyama, T. Takahashi, K. Komatani, T. Ogata, and H. Okuno, “Musical sound separation and synthesis using

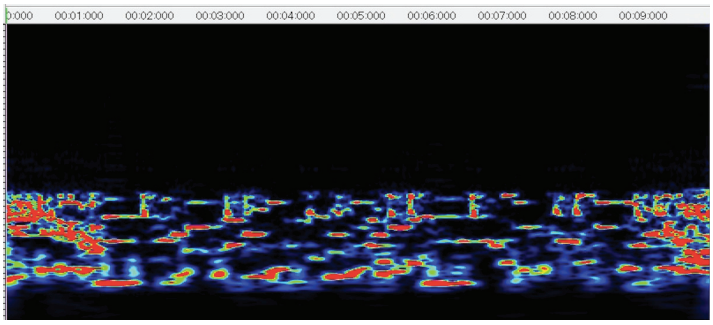

Fig. 5. Spectrogram of the Separated Bass

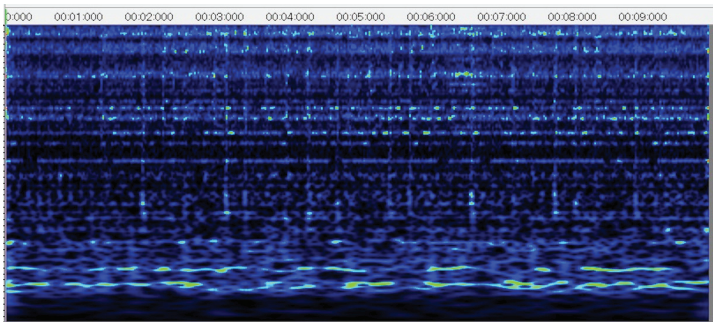

Fig. 6. Spectrogram of the Separated Drums

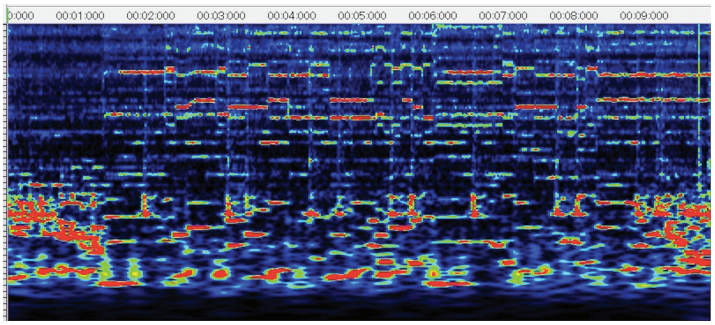

Fig. 7. Spectrogram of the Mixed Down Data of Figures 3 $\sim$ 6

harmonic/inharmonic GMM and NMF for phrase replacing system," *Journal of Information Processing Society of Japan*, vol. 52, no. 12, pp. 3839–3852, 2011, (in Japanese).

[3] M. Nakano, Y. Kitano, J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Polyphonic music signal analysis using non-negative matrix factorization with deformable bases," *IPSJ Special Interest Group on Music and Computer*, vol. 2010-MUS-84, no. 10, pp. 1–6, 2010, (in Japanese).

[4] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[5] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. Okuno, "Constrained parameter estimation of harmonic and inharmonic models for separating polyphonic musical audio signals," *IPSJ Special Interest Group on Music and Computer*, vol. 2007-MUS-37, pp. 81–88, 2007, (in Japanese).

[6] Y. Uchiyama, K. Miyamono, T. Nishimoto, N. Ono, and S. Sagayama, "Automatic chord detection using harmonic/percussive sound separation from music acoustic signals," *IPSJ Special Interest Group on Music and Computer*, vol. 2008-MUS-76, no. 23, pp. 137–142, 2008, (in Japanese).

[7] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. ISMIR*, 2007, pp. 381–386.

[8] M. N. Schmidt and M. Mrup, "Nonnegative matrix factor 2-d deconvolution for blind single channel source separation," in *Proc. ICA*, 2006, pp. 700–707.

[9] H. Kameoka, "Non-negative matrix factorization and its applications to audio signal processing," *Acoustical Science and Technology*, vol. 68, no. 11, pp. 559–565, 2012.

[10] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[11] R. Azuma, "The God of Melodicspeedmetal," http://metaldtm.com/.