PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

A Study on Separation Method Combined Gamma-Process Non-negative Matrix Factorization and Deep Learning.

Satoru JOMAE[1]; Kenko OTA[1]; Hideaki YOSHINO[1]

[1] Nippon Institute of Technology, Japan

ABSTRACT

Accurate analysis of fundamental frequency and chord constitutive notes is a hard problem. However, to solve this problem is important for similar music retrieval and music arrangement etc. Development of an accurate method for sound source separation is required to analyze the fundamental frequency etc. accurately. In this research, we propose a method of sound source separation that combines gamma-process non-negative matrix factorization (GaP-NMF) and Deep Neural Network (DNN). In the proposed method, we first estimate the basis with GaP-NMF. Then, DNN classifies the estimated basis according to musical instruments. The basis estimated by GaP-NMF is emphasized by multiplying with the spectrum template of musical instruments which is specified by DNN. We conducted a sound source separation experiment to verify the performance of the proposed method. Sound sources are composed of multiple musical instrument sounds. As a result of separating a single musical instrument sound from the sound sources, we confirmed that the proposed method improved the SNR by 1.1 dB over the conventional method depending on the data.

Keywords: Sound Source Separation, Nonnegative Matrix Factorization, Deep Learning

1. INTRODUCTION

Currently, we can enjoy the music in various ways due to the progress of Internet technology and Information technology, etc. However, there are some problems in order for us to enjoy music freely. When we play music, we need a musical score. If we cannot obtain the musical score, we need to transcribe the musical score. This is a hard problem for novices. In addition, when we remix existing music, we need a part-by-part recording sound. It is rare that a part-by-part recording sound is provided. Also, when we use a music distribution service, we would like to retrieve music similar to the music which we like. We need to understand the musical information such as pitch, tempo, and melody so as to realize the similar music retrieval.

In fact, many researchers have carried out research on pitch and melody estimation. For example, Zhang et al. have reported melody estimation using a particle filter (1). However, when the sound data consists of multiple sound sources, the estimation performance of this method is low.

Hence, a sound source separation technique is required to solve these above-mentioned problems. Sound source separation techniques with Non-negative matrix factorization (NMF) (2) have been proposed to solve the above-mentioned problems. When we utilize NMF, we need to specify the number of bases in advance. The Gamma Process Non-negative Matrix Factorization (GaP-NMF)(3) has solved the problem of NMF by introducing Gamma process to NMF and by extending NMF by Bayesian nonparametric method. GaP-NMF makes it possible to separate while inferring the number of unknown sound sources. However, there is also a problem with GaP-NMF. We expect to obtain a basis matrix as a set of the spectrum of a single tone so as to realize the multiple sound source separation. The basis matrix which is estimated by GaP-NMF is not a set of the spectrum of a single tone. Hence, in this study, we propose a method that introduces deep learning to the learning process of a basis matrix of GaP-NMF.

In this paper, Section 2 describes related research and the proposed method. Section 3 describes the test data and evaluation method used in the evaluation experiment. Section 4 describes the examination of the results of evaluation experiments, and Section 5 shows the conclusions of this study.

2. METHODS

In this section, we introduce our proposed method. Our proposed method consists of GaP-NMF (Gamma Process Non-negative Matrix Factorization and deep learning.

2.1 Non-negative Matrix Factorization (NMF)(2)

NMF is a method for factorizing an input data matrix $\boldsymbol{X} \in \mathbb{R}_{+}^{m \times n}$ into two smaller output matrices $\boldsymbol{W} \in \mathbb{R}_{+}^{m \times k}$ and $\boldsymbol{H} \in \mathbb{R}_{+}^{k \times n}$. When applying NMF to a musical acoustic signal, we use a sound spectrogram as an input matrix $\boldsymbol{X}$. In the separation process, we assume that the matrix X consists of k bases. One output matrix $\boldsymbol{W}$ represents the basis matrix of a sound source. The other output matrix $\boldsymbol{H}$ represents the activation matrix of each basis.

The problem with NMF is that we need to specify the number of basis in advance. In addition, we expect to obtain a basis matrix as a set of the spectrum of a single tone, but the basis matrix which is estimated by the basic NMF is not a set of the spectrum of a single tone.

2.2 Gamma Process Non-negative Matrix Factorization (GaP-NMF)(3)

When applying NMF to a musical acoustic sound, we need to specify the number of basis in advance. However, when actually applying NMF to the musical acoustic sound, it is not always possible to specify the number of bases to be estimated. Hence, GaP-NMF has been introduced a nonparametric Bayesian estimation method to NMF to solve the above-mentioned problem. GaP-NMF estimates the basis matrix and the activation matrix of the musical acoustic signal by updating the parameters for a probability distribution. And, GaP-NMF estimates the number of bases simultaneously. We can obtain a basis matrix whose number of bases is restricted not to use more bases than necessary. However, the basis matrix which is estimated by GaP-NMF is also not a set of the spectrum of a single tone. The reason for this problem is GaP-NMF estimates a basis matrix by minimizing the error with an observation matrix. Hence, we need to add a constraint to parameter update process so as to estimate a set of the spectrum of a single tone as the basis matrix. We call GaP-NMF as the conventional method in this paper.

2.3 Proposed method

Figure 1 shows the processing flow of our proposed method. Our proposed method makes template database which is a set of the spectrum of a single tone in advance and classifies each estimated basis into a template spectrum. The estimated basis is emphasized by averaging with the template spectrum to approximate each estimated basis to the spectrum of a single tone. In our proposed method, we utilize GaP-NMF so as to estimate the basis matrix and the activation matrix, and we utilize Convolutional Neural Network (CNN) so as to classify the basis into a template spectrum.

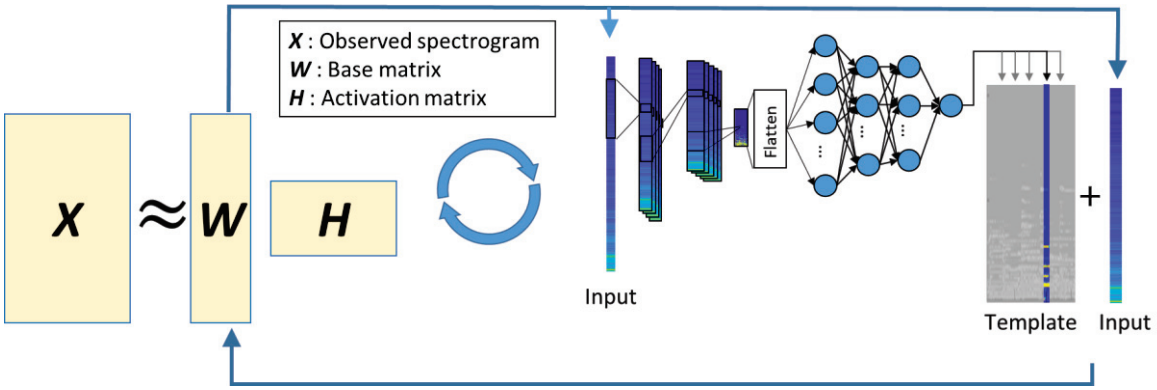


Fig. 1 The processing flow of our proposed method

Table 1 Parameters of the structure of CNN

No.	Type	Description
1	Input	1025x1x1 images with 'zerocenter' normalization
2	Convolution	48 4x1x1 convolutions with stride [1 1] and padding [0 0 0 0]
3	ReLU	ReLU
4	Batch Normalization	Batch normalization with 48 channels
5	Max Pooling	2x1 max pooling with stride [2 1] and padding [0 0 0 0]
6	Convolution	256 2x1x48 convolutions with stride [1 1] and padding 'same'
7	ReLU	ReLU
8	Batch Normalization	Batch normalization with 256 channels
9	Max Pooling	2x1 max pooling with stride [2 1] and padding [0 0 0 0]
10	Convolution	384 3x1x256 convolutions with stride [2 1] and padding [1 1 1 1]
11	ReLU	ReLU
12	Convolution	384 3x1x384 convolutions with stride [2 1] and padding [1 1 1 1]
13	ReLU	ReLU
14	Convolution	256 3x1x384 convolutions with stride [2 1] and padding [1 1 1 1]
15	ReLU	ReLU
16	Max Pooling	2x1 max pooling with stride [2 1] and padding [0 0 0 0]
17	Fully Connected	1025 fully connected layer
18	ReLU	ReLU
19	Dropout	50% dropout
20	Fully Connected	1025 fully connected layer
21	ReLU	ReLU
22	Dropout	50% dropout
23	Fully Connected	3 fully connected layer
24	Softmax	softmax
25	Classification Output	

Table 2 Parameters of the training of CNN

Option	Value
Solver	SGDM
InitialLearnRate	0.001
MaxEpochs	20
MiniBatchSize	40
Shuffle	every-epoch

3. EXPERIMENT

We performed source separation experiments so as to confirm the performance of the proposed method.

3.1 Conditions and methods

We utilized 200 notes of each musical instrument, piano, vibraphone, and bass guitar from "RWC Research Music Database: Musical Instruments(4)" as the training data of CNN. Tables 1 and 2 show the parameters of the structure of CNN and the training of CNN.

We utilized jazz music from the database "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research(5)" for acoustic research as the test data. Both the training data and the test data were sampling frequency of 44,100 Hz and quantization bit rate of 16 bits. Test data consists of piano, percussion, bass guitar, trumpet, and saxophone. Since the test data was recorded by stereo, we converted the stereo data to the monaural data. The duration of the test data was about 90 seconds, and we separated the test data into the short part data whose duration is 10 seconds. As the pre-processing of sound source separation, we converted the test data to the power spectrogram using short time Fourier transform. We utilized Hamming window for Fast Fourier transform whose frame length was 2048 points, and frame shift was 1024 points. Because we thought that it would be better if sufficient resolution could be obtained to separate the tone of C3 and higher.

In this experiment, the maximum number of bases was set to 30 in the conventional method and the proposed method, and the number of updates of each output matrix was set to 30. The basis spectrum was enhanced by the convolution network on the frequency region once every three updates of GaP-NMF.

We utilized S/N ratio given by the following equation as the index of separation performance.

$$\text{S/N ratio} = \mathbf{10}\,\mathbf{log_{10}}\frac{\sum_t \boldsymbol{s(t)^2}}{\sum_t \big(\boldsymbol{s(t)} - \boldsymbol{n(t)}\big)^2}\ \mathbf{[dB]} \tag{1}$$

Here, $\boldsymbol{s(t)}$ is the original signal before mixing, and $\boldsymbol{n(t)}$ is the estimated signal after separation. In this research, we assumed the acoustic signal of the piano as the original signal.

3.2 RESULT

Figure 2 shows the separation accuracy of the conventional method and the proposed method. The S/N ratio improved in most parts. The S/N ratio improved significantly in Parts 1, 2 and 7-9. In Part 2, in particular, significant improvement in accuracy was confirmed compared to the conventional method. Figure 3 shows the input spectrogram of Part 2. There were few overlapping sounds in the low-frequency range, and there were few variations in pitch. Also, no significant difference in performance was found in Parts 3 and 6. In Part 4, the conventional method was higher separation accuracy than the proposed method. Figure 4 shows the input spectrogram of Part 4. This part was rich in the change of pitch and was a sound source with many occurrences of sound.

Figures 5 and 6 show the basis matrices of Part 2 and 4, respectively. The basis matrix of the conventional method did not contain the high-frequency component. On the other hand, the basis matrix of the proposed method can be possible to represent the high-frequency component. In addition, the basis matrix of the conventional method was sparser than that of the proposed method.

Figures 7 and 8 show the spectrogram of Part 2 estimated by the conventional method and the proposed method, respectively. Figures 9 and 10 show the spectrogram of Part 4 estimated by the conventional method and the proposed method, respectively. The spectrogram estimated by the conventional method does not contain high-frequency components as well as the basis matrix. Especially in the output spectrogram of part 4 by the proposed method shown in Fig. 10, high-frequency components are greatly emphasized.

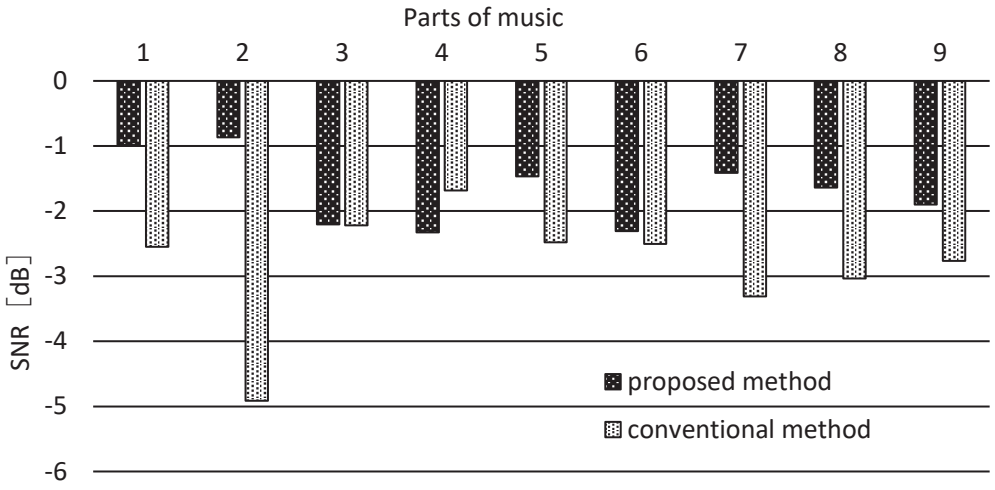


Fig. 2 Separation accuracy

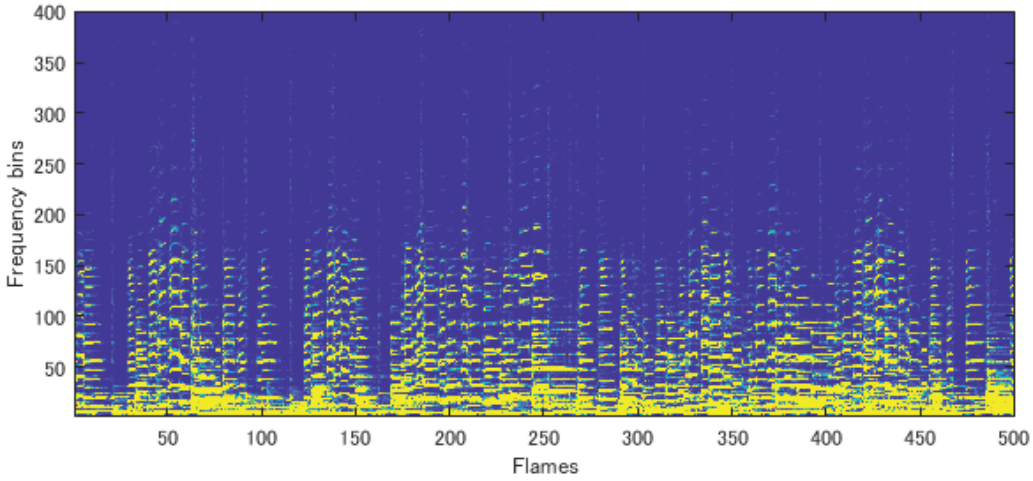


Fig. 3 Input spectrogram of Part 2

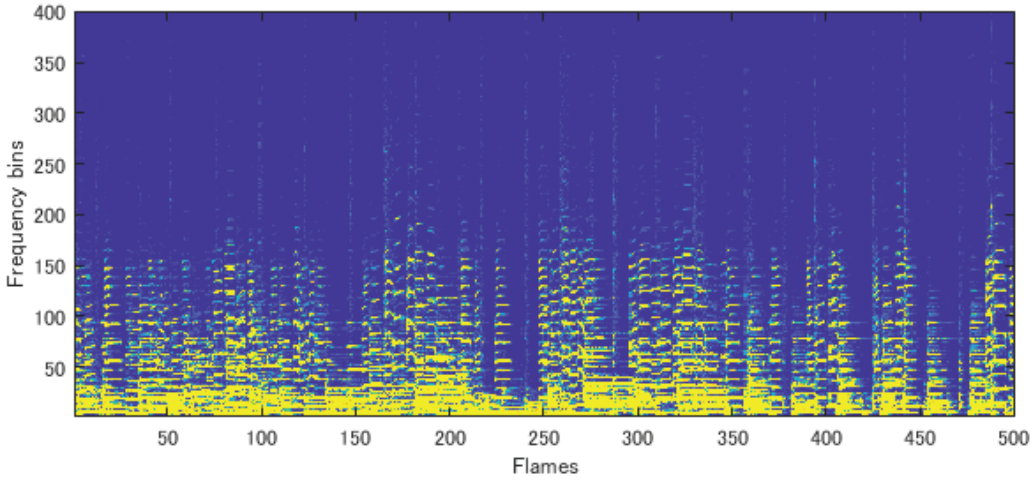
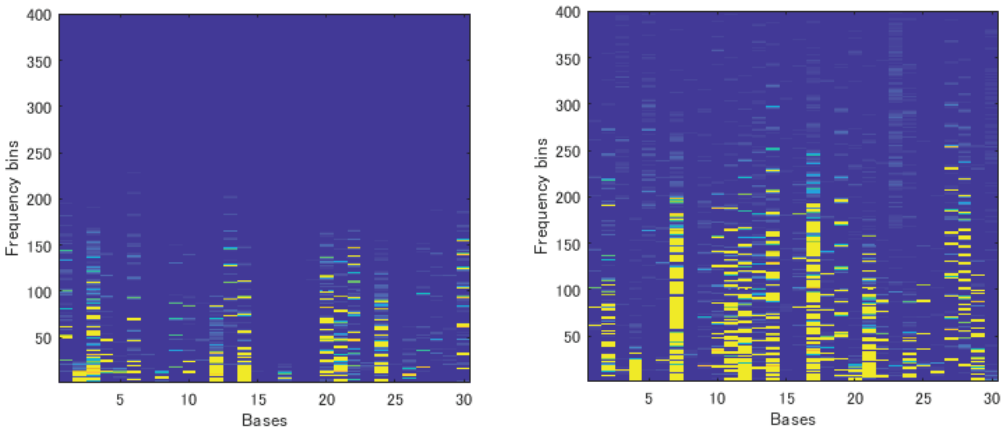
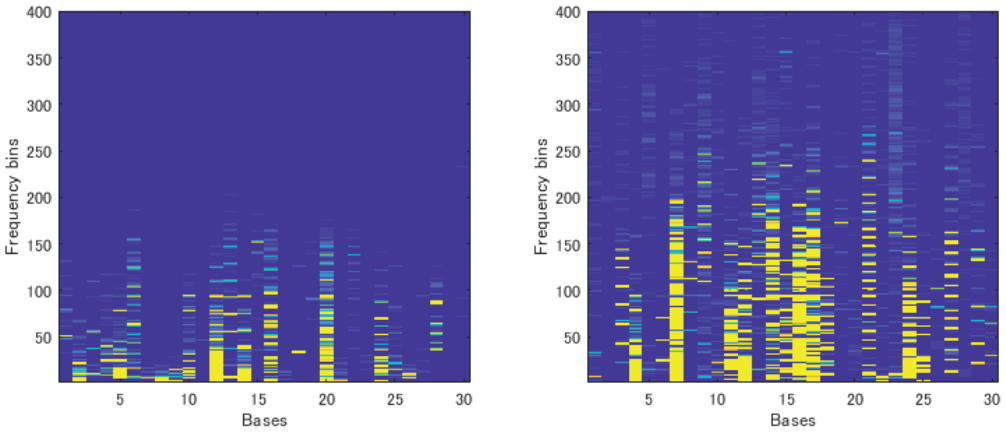


Fig. 4 Input spectrogram of Part 4



(a) Conventional method (b) Proposed method

Fig.5 Estimated basis matrices of Part 2



(a) Conventional method (b) Proposed method

Fig.6 Estimated basis matrices of Part 4

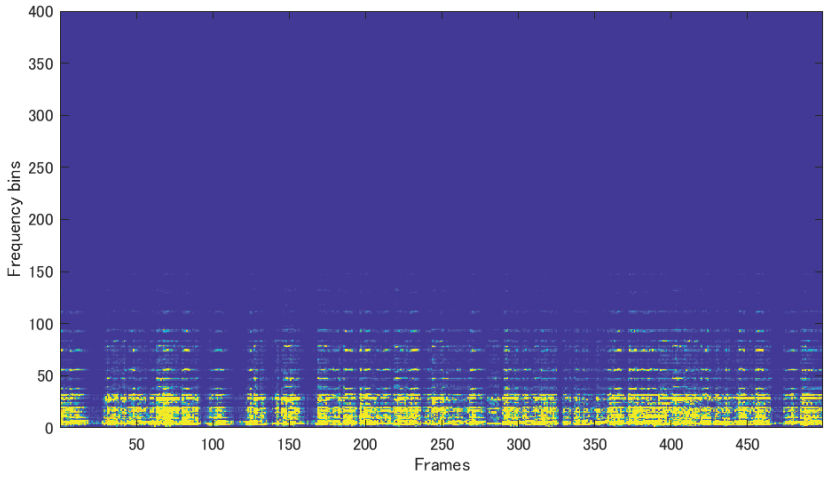


Fig. 7 Spectrogram of Part 2 estimated by the conventional method

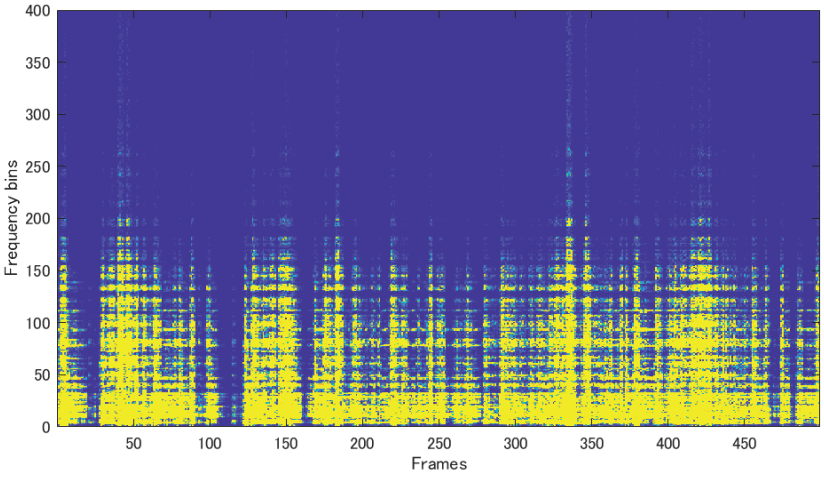


Fig. 8 Spectrogram of Part 2 estimated by the proposed method

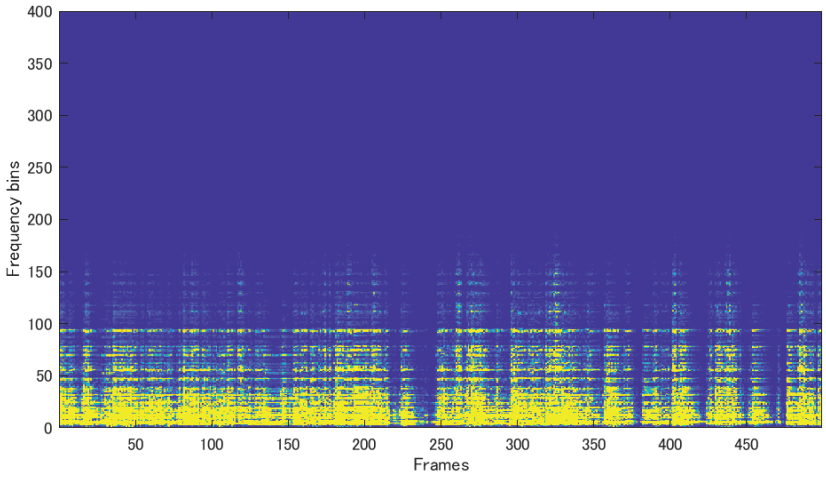


Fig. 9 Spectrogram of Part 4 estimated by the conventional method

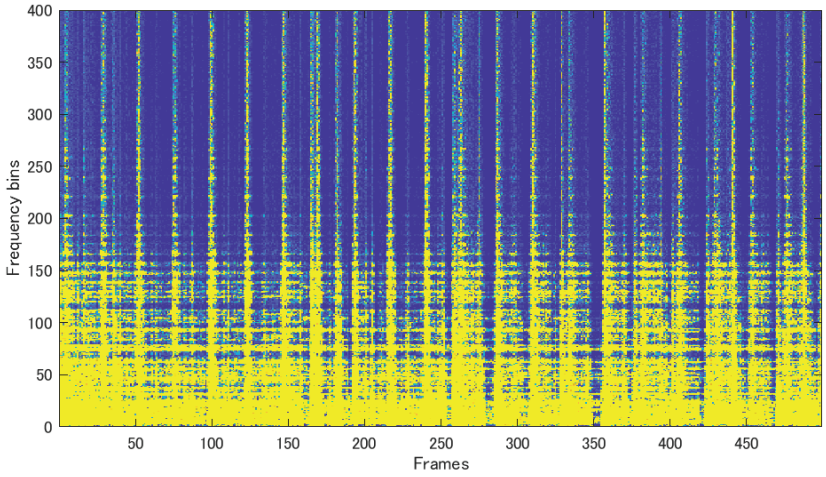


Fig. 10 Spectrogram of Part 4 estimated by the proposed method

4. DISCUSSION

The experimental results show that the proposed method is superior to the conventional method in many parts. The S/N ratio of the conventional method is superior to the proposed method in a part. Especially when comparing the part where the S/N ratio improved and the part where it did not improve, there were differences in the number of variations in pitch. The proposed method is advantageous for the data with relatively slow change in pitch compared to the conventional method. However, for sound data rich in pitch change, the conventional method showed high separation accuracy. Sound data rich in pitch change contains many sound sources whose duration is short. From this, in the proposed method, it can be inferred that the feature of the basis vector corresponding to the sound source with a small generation time or the number of pronunciations is caused to disappear by the enhancement of the basis matrix by deep learning.

The disappearance of the feature of the basis vector is due to the emphasis by deep learning with the basis matrix not fully converged. This is due to the fact that basis vectors with similar pitches could not be determined sufficiently by deep learning. As a result, sounds with similar pitches were grouped together and it was not possible to express the sound of the separation object sufficiently.

5. CONCLUSIONS

In this study, we extended GaP-NMF using deep learning method. The test data this time gave superior results to the conventional method. In particular, significant improvement in the S／N ratio was confirmed for sounds with a gradual change in pitch. However, compared to the conventional method, significant improvement in separation accuracy was not obtained for sounds rich in change in pitch and sounds with a large number of pronunciations. It can be inferred that this is because the feature of the sound source with a short pronunciation time disappears by emphasizing the spectrum by deep learning. In the future, it is necessary to consider to solve this problem.

REFERENCES

1. W. Zhang, Chen Z, Yin F, Zhang Q. Melody Extraction From Polyphonic Music Using Particle Filter and Dynamic Programming. IEEEACM Trans Audio Speech Lang Process. 2018 Sep;26(9):1620–32.
2. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999 Oct;401(6755):788–91.
3. Hoffman MD, Blei DM, Cook PR. Bayesian Nonparametric Matrix Factorization for Recorded Music. In: Proceedings of the 27th International Conference on International Conference on Machine Learning [Internet]. USA: Omnipress; 2010 [cited 2018 Dec 25]. p. 439–446. (ICML'10). Available from: http://dl.acm.org/citation.cfm?id=3104322.3104379
4. Goto M. Development of the RWC music database. Proc 18th Int Congr Acoust ICA 2004. 2004 May;Vol. 1:553–6.
5. Bittner R, Salamon J, Tierney M, Mauch M, Cannam C, Bello J. MedleyDB: A MULTITRACK DATASET FOR ANNOTATION-INTENSIVE MIR RESEARCH. 2014;6.