Vocal Separation by Constrained Non-Negative Matrix Factorization

Eri Ochiai*, Takanori Fujisawa†, Masaaki Ikehara‡
EEE Dept., Keio Univ., Yokohama, Kanagawa 223-8522, Japan
E-mail:{ochiai*, fujisawa†, ikehara‡}@tkhm.elec.keio.ac.jp

***Abstract*—The vocal separation is to separate vocal part and remove the accompaniment part from the mixed music data. Vocal part include many information, singer, lyric and emotion of the song. If we can extraction only the vocal part from the original sound from CD source, it can be applied to various applications. In this paper, we propose a new method to take out the natural vocal parts from mixed music by using non-negative matrix factorization (NMF). This NMF-based framework separates the harmonic, percussive, and vocal structures from the input signal. We impose the constraint into each component to enforce its feature such as harmonic or temporal continuity. In addition, we propose a framework utilizing the prior information in order to achieve the valid vocal separation over this mathematical procedure. The experiments over some vocal databases show the proposed framework has the superior separation performance compared to the conventional methods. Also by considering the characteristics of music, we intend to obtain high accuracy result.**

I. Introduction

Many approaches have been proposed [1]–[5]. We can extraction only vocal from the original mixed CD sound source, it can be applied various applications, such as vocal score creation, lyric recognition, music information retrieval and so on. Among the vocal separation of acoustic signals, there are processing methods using statistical model [2], [3] and using the music feature [4], [5]. The estimation method utilizing a statistical model needs to learn the vocal fundamental frequency f_0 of the accompaniment. However estimation f_0 is difficult, because the results may have a big distortion due to accompaniment. The other methods utilize the feature of vocal signals. Music signal consists of the vocal and the accompaniment parts. According to the conventional model in [6], the accompaniment part is represented by the mixture of harmonic and percussive structures. And vocal parts include many information such as pitch, formant and so on. Therefore conventional methods aim to separate vocal part by emphasizing the features of each part.

One of the popular method for musical source segmentation is non-negative matrix factorization (NMF) [7]. NMF-based vocal separation has been proposed by [8], it utilizes the spectrogram of the input signal by modeling itself using machine learning techniques from training data generated from prior information. This makes it possible to obtain a fundamental character of the input matrix. In this paper, we extend the NMF-based vocal separation framework by introducing the harmonic-percussive-vocal music model. The NMF with constraint achieves to separate the input spectrogram while keeping the feature of each part. The experiments on the vocal database show that the proposed framework performs better separation results compared to the conventional methods.

II. Conventional Approach of Vocal Separation

The music analysis by NMF needs a time-frequency representation of music signal $\mathbf{X}$. The time domains are indexed by t as a frame number and the frequency domain are indexed by f. This representation is obtained by short-time fourier transform (STFT). And the time-frequency representation of music signal can be reconstructed by inverse STFT.

A. Music Signal Model of Jeong Method

Music signal is assumed to be the mixture of vocal and accompaniment parts. The accompaniment part can be separated into harmonic and percussive structures. The harmonic component shapes the spectrogram with the smoothed envelopes over periods of time. On the other hand, the percussive component produces the impulsive ridges on the frequency domain of the spectrogram. The harmonic and percussive structures have characteristics of time continuity and frequency continuity respectively. Jeong et al. proposed a music signal model based on harmonic, percussive and vocal structure [6]. Original music signal $\mathbf{X}$ can be represented by the sum of these structures, as follows:

$$\mathbf{X} \approx \mathbf{Y} = \mathbf{Y}_{\mathrm{H}} + \mathbf{Y}_{\mathrm{P}} + \mathbf{Y}_{\mathrm{V}} \tag{1}$$

$\mathbf{X}$ denotes the original music signal and $\mathbf{Y}$ denotes signal model. $\mathbf{Y}_{\mathrm{H}}, \mathbf{Y}_{\mathrm{P}}, \mathbf{Y}_{\mathrm{V}}$ denotes the spectrograms representing the harmonic, the percussive and the vocal component respectively and the mixture of these three components approximates the input signal as $\mathbf{Y}$. Harmonic and percussive structure can be obtained by minimizing the penalties shown in (2) and (3).

$$C_{\mathrm{H}} = \frac{\lambda_{\mathrm{H}}}{2} \sum_{f,t} \left(Y_{\mathrm{H},f,t-1} - Y_{\mathrm{H},f,t}\right)^2 \tag{2}$$

$$C_{\mathrm{P}} = \frac{\lambda_{\mathrm{P}}}{2} \sum_{f,t} \left(Y_{\mathrm{P},f-1,t} - Y_{\mathrm{P},f,t}\right)^2 \tag{3}$$

λ_{H} and λ_{P} is weight. The penalty C_{H} enforces the temporal smoothness by reducing the differences between neighbored elements on the time domain. And the penalty C_{P} ensures the vertical ridges over the frequency domain. This method estimates $\mathbf{Y}_{\mathrm{H}}$ and $\mathbf{Y}_{\mathrm{P}}$ from $\mathbf{Y}$ and obtains the vocal component $\mathbf{X}_{\mathrm{V}}$ as residual.

B. NMF Algorithm with β-divergence and MU Update Rule

In this paper, we reformulate this Jeong's model by NMF-based decomposition framework. NMF is a mathematical technique decomposing a non-negative matrix $\mathbf{X}$ into two non-negative matrices $\mathbf{S}$ and $\mathbf{A}$ [7]. In music signal analysis, matrix $\mathbf{X}$ is time-frequency spectrogram obtained by STFT. $\mathbf{S}$ is called

the base term which contains a set of basis spectra. It forms the spectrogram $\mathbf{X}$ by multiplying with matrix $\mathbf{A}$, which is called the activity term. The expressed NMF is

$$\mathbf{X} \approx \mathbf{Y} = \mathbf{S}\mathbf{A}. \tag{4}$$

Estimating $\mathbf{S}$ and $\mathbf{A}$, first we initialize them by random number. And NMF decomposes $\mathbf{X}$ into $\mathbf{S}$ and $\mathbf{A}$ by minimizing β-divergence [9] represented by following equation (5):

$$d(\mathbf{X}|\mathbf{Y}) = \frac{1}{\beta(\beta-1)} \sum_{f,t} \left| X_{f,t}^{\beta} + (\beta-1)Y_{f,t}^{\beta} - \beta X_{f,t} Y_{f,t}^{(\beta-1)} \right| \tag{5}$$

The commonly used approach to achieve this optimization is "multiplicative update method" (MU rule) [10]. MU rule minimizes the distortion $d(\mathbf{X}|\mathbf{Y})$ by mutually applying the following equations (6) to the component $\mathbf{S}$ and $\mathbf{A}$:

$$\mathbf{S} \leftarrow \mathbf{S} \circ \frac{\nabla_{\mathbf{S}}^{-} d(\mathbf{X}|\mathbf{Y})}{\nabla_{\mathbf{S}}^{+} d(\mathbf{X}|\mathbf{Y})},\ \mathbf{A} \leftarrow \mathbf{A} \circ \frac{\nabla_{\mathbf{A}}^{-} d(\mathbf{X}|\mathbf{Y})}{\nabla_{\mathbf{A}}^{+} d(\mathbf{X}|\mathbf{Y})} \tag{6}$$
$$\left(\nabla_{\mathbf{S}} d = \nabla_{\mathbf{S}}^{+} d - \nabla_{\mathbf{S}}^{-} d,\ \nabla_{\mathbf{S}}^{+} d \geq 0, \nabla_{\mathbf{S}}^{-} d \geq 0\right)$$

The operation $\circ$ denotes the element-wise multiplication. The division of two matrices are also conducted element-wise. We calculate positive and negative terms ∇^{+}, ∇^{-} dividing the gradient ∇d. In the case of minimizing the β-divergence, the update rules are represented by following equations (7):

$$\mathbf{S} \leftarrow \mathbf{S} \circ \frac{(\mathbf{Y}^{\beta-2} \circ \mathbf{X})\mathbf{A}^{T}}{\mathbf{Y}^{\beta-1}\mathbf{A}^{T}},\ \mathbf{A} \leftarrow \mathbf{A} \circ \frac{\mathbf{S}^{T}(\mathbf{Y}^{\beta-2} \circ \mathbf{X})}{\mathbf{S}^{T}\mathbf{Y}^{\beta-1}} \tag{7}$$

III. The Proposed Vocal Separation with Pre-Learned NMF

We aim at improving the pre-learned NMF model [11] by introducing Jeong's penalty terms (9), (10), (11). And we newly add vocal part for temporary characteristic constraint.

A. Proposed Model Reformulating Jeong's Model by NMF

From equations (1) and (4), we reformulate the model of composing music spectrogram in suitable form for NMF as follows:

$$\mathbf{X} \approx \mathbf{Y} = \mathbf{S}_{\mathrm{H}}\mathbf{A}_{\mathrm{H}} + \mathbf{S}_{\mathrm{P}}\mathbf{A}_{\mathrm{P}} + \mathbf{S}_{\mathrm{V}}\mathbf{A}_{\mathrm{V}} \tag{8}$$

$\mathbf{S}_{\mathrm{H}}$ denotes a harmonic base, $\mathbf{A}_{\mathrm{H}}$ denotes a harmonic activity and $\mathbf{S}_{\mathrm{P}}$ denotes a percussive base, $\mathbf{A}_{\mathrm{P}}$ denotes a percussive activity and $\mathbf{S}_{\mathrm{V}}$ denotes a vocal base, $\mathbf{A}_{\mathrm{V}}$ denotes a vocal activity. In NMF, a harmonic constraint condition is imposed to $\mathbf{A}_{\mathrm{H}}$ that include temporal information. As well, a percussive constraint condition is imposed to $\mathbf{S}_{\mathrm{P}}$ that include frequency information. We newly impose a vocal constraint condition to $\mathbf{A}_{\mathrm{V}}$ because we regard a vocal part as a temporary. So we define constraint conditions as follows

$$C_H(\mathbf{A}) = \frac{\lambda_{\mathrm{H}}}{2} \sum_{i,t} \left(A_{\mathrm{H},i,t-1} - A_{\mathrm{H},i,t}\right)^2 \tag{9}$$

$$C_P(\mathbf{S}) = \frac{\lambda_{\mathrm{P}}}{2} \sum_{f,i} \left(S_{\mathrm{P},f-1,i} - S_{\mathrm{P},f,i}\right)^2 \tag{10}$$

$$C_V(\mathbf{A}) = \frac{\lambda_{\mathrm{V}}}{2} \sum_{i,t} \left(A_{\mathrm{V},i,t-1} - A_{\mathrm{V},i,t}\right)^2 \tag{11}$$

Based on these constraints (9) to (11), we redefine the vocal separation problem to the equation $d'(\mathbf{X}|\mathbf{Y})$ as follow.

$$d'(\mathbf{X}|\mathbf{Y}) = d(\mathbf{X}|\mathbf{Y}) + \lambda_{\mathrm{H}} C_{\mathrm{H}}(\mathbf{A}) + \lambda_{\mathrm{P}} C_{\mathrm{P}}(\mathbf{S}) + \lambda_{\mathrm{V}} C_{\mathrm{V}}(\mathbf{A}) \tag{12}$$

By utilizing the MU rule, the update rules of each component is derived as the equation (13) to (18).

$$\mathbf{S}_{\mathrm{H}} \leftarrow \mathbf{S}_{\mathrm{H}} \circ \frac{(\mathbf{Y}^{\beta-2} \circ \mathbf{X})\mathbf{A}_{\mathrm{H}}^{T}}{\mathbf{Y}^{\beta-1}\mathbf{A}_{\mathrm{H}}^{T}} \tag{13}$$

$$\mathbf{A}_{\mathrm{H}} \leftarrow \mathbf{A}_{\mathrm{H}} \circ \frac{\mathbf{S}_{\mathrm{H}}^{T}(\mathbf{Y}^{\beta-2} \circ \mathbf{X}) + \lambda_{\mathrm{H}}[A_{\mathrm{H},i,t-1} + A_{\mathrm{H},i,t+1}]}{\mathbf{S}_{\mathrm{H}}^{T}\mathbf{Y}^{\beta-1} + 2\lambda_{\mathrm{H}}\mathbf{A}_{\mathrm{H}}} \tag{14}$$

$$\mathbf{S}_{\mathrm{P}} \leftarrow \mathbf{S}_{\mathrm{P}} \circ \frac{(\mathbf{Y}^{\beta-2} \circ \mathbf{X})\mathbf{A}_{\mathrm{P}}^{T} + \lambda_{\mathrm{P}}[S_{\mathrm{P},f-1,i} + S_{\mathrm{P},f+1,i}]}{\mathbf{Y}^{\beta-1}\mathbf{A}_{\mathrm{P}}^{T} + 2\lambda_{\mathrm{P}}\mathbf{S}_{\mathrm{P}}} \tag{15}$$

$$\mathbf{A}_{\mathrm{P}} \leftarrow \mathbf{A}_{\mathrm{P}} \circ \frac{\mathbf{S}_{\mathrm{P}}^{T}(\mathbf{Y}^{\beta-2} \circ \mathbf{X})}{\mathbf{S}_{\mathrm{P}}^{T}\mathbf{Y}^{\beta-1}} \tag{16}$$

$$\mathbf{S}_{\mathrm{V}} \leftarrow \mathbf{S}_{\mathrm{V}} \circ \frac{(\mathbf{Y}^{\beta-2} \circ \mathbf{X})\mathbf{A}_{\mathrm{V}}^{T}}{\mathbf{Y}^{\beta-1}\mathbf{A}_{\mathrm{V}}^{T}} \tag{17}$$

$$\mathbf{A}_{\mathrm{V}} \leftarrow \mathbf{A}_{\mathrm{V}} \circ \frac{\mathbf{S}_{\mathrm{V}}^{T}(\mathbf{Y}^{\beta-2} \circ \mathbf{X}) + \lambda_{\mathrm{V}}[A_{\mathrm{V},i,t-1} + A_{\mathrm{V},i,t+1}]}{\mathbf{S}_{\mathrm{V}}^{T}\mathbf{Y}^{\beta-1} + 2\lambda_{\mathrm{V}}\mathbf{A}_{\mathrm{V}}} \tag{18}$$

The member enclosed by brackets means the matrix composed of the elements inside the brackets.

B. Determining Harmonic and Percussive Initial Values for NMF calculation

In the proposed method, the framework for separating vocal part is shown in Figure 1. To estimate a music signal, we need to initialize the values of $\mathbf{S}$ and $\mathbf{A}$ by some prior information. As a prior information, we prepare the signal consisting of only accompaniment part. First we describe method to initialize all $\mathbf{S}_{\mathrm{P}}$ and $\mathbf{S}_{\mathrm{H}}$. We bring the number of spectra at random from the spectrogram of prior information. Then we arrange these spectra by the energy order, and make larger energy spectra part for the components of the percussive base, as well smaller energy spectra part for the components of the harmonic base.

We to create vocal base in three methods, a overtone structure, a overtone structure shaped in formant, and a sample vowel spectrogram. In the first method, we manually form the spectrogram of harmonic vocal basis of each musical pitch. The real vocal spectrogram includes more information in low frequency and less information in high frequency. So we create the overtone structure where the low frequency gains are strong and the high frequency gains are weak. The overtone model is also used in [12]. Second the overtone structure shaped by formant made by reducing of the overtone structure that we made. The formant structure is obtained by cepstrum method [13]. Third we generated the spectra from the sample vowel data obtained from the database of [14]. $\mathbf{A}_{\mathrm{P}}$, $\mathbf{A}_{\mathrm{H}}$ and $\mathbf{A}_{\mathrm{V}}$ are initialized with a random number.

C. Learning step and Estimating step of the Proposed Method

The proposed method consists of three steps as shown in Fig1. The first step is learning step. The second step is estimation step. The third step is post-processing step.

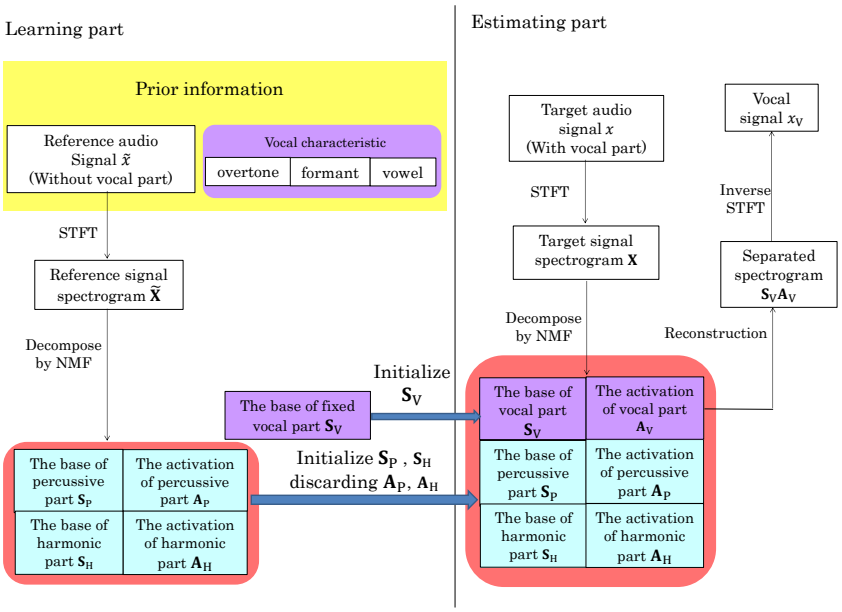


Fig. 1. The framework of proposed NMF method

Algorithm 1 The procedure of the proposed method

Learning Step

1: STFT prior information signal to obtain input spectrogram $\tilde{\mathbf{X}}$.
2: Randomly choose spectrum from only accompaniment music.
3: Rearrange spectra in order to energy.
4: Initialize $\mathbf{S}_P$ and $\mathbf{S}_H$.
5: Initialize $\mathbf{S}_V$ by overtone structure or formant or vowel sample.
6: **for** some iteration **do**
7: Update the initial part $\mathbf{S}_H$ and $\mathbf{S}_P$ respectively as (13) to (16).
8: **end for**

Estimating Step

1: Transform the target music into spectrogram $\mathbf{X}$ by STFT.
2: **for** Some iteration **do**
3: Update the basis part $\mathbf{S}_H$, $\mathbf{S}_P$, $\mathbf{S}_V$ respectively as (13) to (18).
4: **end for**
5: Reconstruct vocal part $\boldsymbol{x}_V$ by applying inverse STFT to $\mathbf{S}_V\mathbf{A}_V$.

The learning step is to produce the initial value of $\mathbf{S}_P$ and $\mathbf{S}_H$ from prior information. It extracts the basis $\mathbf{S}_P$ and $\mathbf{S}_H$ from the spectrogram of prior information $\tilde{\mathbf{X}}$. The update rules use (13) to (16) in this step. We also obtain the initial value of $\mathbf{S}_V$ as described in the previous section.

By utilizing $\mathbf{S}_P, \mathbf{S}_H, \mathbf{S}_V$ from learning step, the estimation step decomposes the target spectrogram $\mathbf{X}$ into $\mathbf{S}$ and $\mathbf{A}$ parts using the update rules (13) to (18). After estimating all $\mathbf{S}$ and $\mathbf{A}$, we reconstruct vocal part $\mathbf{S}_V\mathbf{A}_V$ by inverse STFT.

As post-processing, we cut-off low vocal frequency of the estimation results because the vocal component does not include low frequency range. Post-processing can improve a separation performance. The pseudo code for learning and estimating step are shown in Algorithm 1 respectively.

IV. Evaluation

A. Data set

To quantitatively evaluate the proposed vocal separation algorithm, we used the MIR-1K database in [15]. The database consists of 1000 music clips sung by amateur singers. The song was sampled by 16-bit precision with the sampling frequency of 16000Hz. The vocal parts and the accompaniment parts are recorded in separate channels. When estimating them, we sum each channel. For STFT analysis, we use the Fourier window size 1024 samples with a 512 overlapping. The parameters are shown in Table I. And the cut-off frequency of post-filtering is set to 1500Hz. This value is available from try and error.

TABLE I
THE PARAMETERS OF WEIGHT λ AND NUMBER OF BASE INDEX FOR EACH VOCAL STRUCTURE IN PROPOSED METHOD.

Parameters	overtone	formant	vowel
λ_H	150	50	80
λ_P	80	200	200
λ_V	0.001	0.0003	0.001
Base index of harmonic	100	60	100
Base index of percussive	100	80	100
Base index of vocal	100	80	102

TABLE II
VOCAL SEPARATION RESULT IN SDR[dB]. JEONG IS CONVENTIONAL METHOD AND NMF IS NMF WITHOUT CONSTRAINT CONDITION. *abjones*, *amy*, *bobon*, *jmzen* IS DIFFERENT CHARACTERISTIC MUSICS AND AVERAGE IS MEAN OF 4MUSICS RESULTS.

Music	*abjones*	*amy*	*bobon*	*jmzen*	Average
Jenog model [11]	2.19	3.79	1.45	1.24	2.17
NMF	2.11	0.31	0.51	1.49	1.11
Proposed method of overtone	3.46	5.48	0.81	2.24	3.00
Proposed method of formant	3.00	5.51	1.63	1.97	3.03
Proposed method of vowel	3.74	5.66	1.83	3.34	3.64

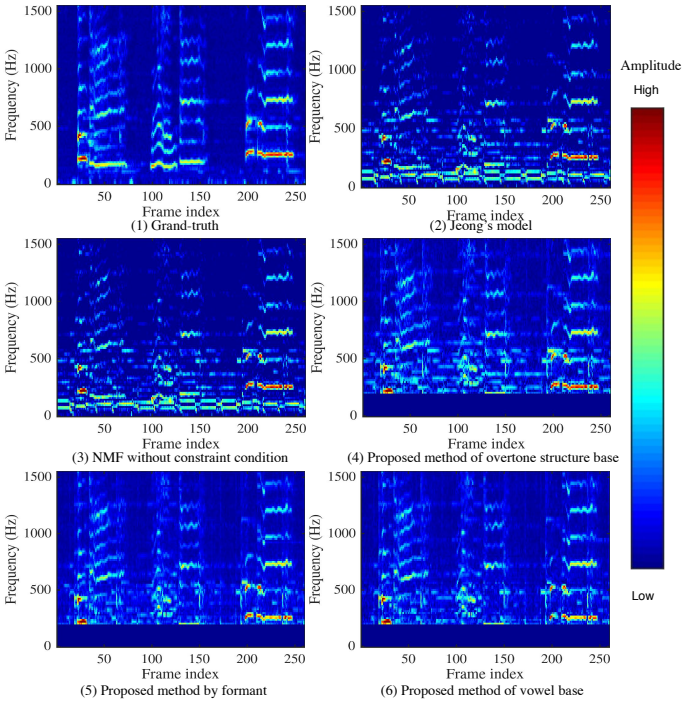
As an evaluation, we use Signal-to-Distortion Ratio(SDR), which is represented by the following equation:

$$\mathrm{SDR} = 10\log_{10}\left(\frac{\sum_t x(t)^2}{\sum_t |y(t)-x(t)|^2}\right)[\mathrm{dB}] \qquad (19)$$

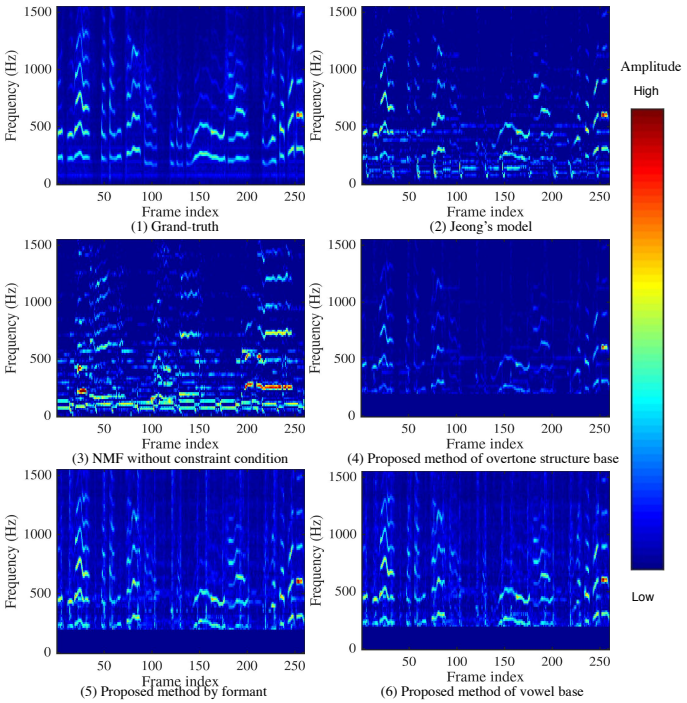
$x(t)$ is grand-truth and $y(t)$ is estimated signal indexed by the time t.

B. Vocal separation

We estimated 4 musics and compared to the conventional methods [7], [11]. We have implemented the proposed method 20 times and we took the average of SDR results. From Table



(a) The extracted vocal spectrogram of "*abjones*"



(b) The extracted vocal spectrogram of "*amy*"

Fig. 2. Spectrogram results.

II, the proposed method has better values than conventional method. The estimation initialized by vowel data performs the best result, however the results of *abjones* and *amy* have estimation error in high frequency part from Fig2. According to these figures, the proposed method can recover the closest spectrogram to the grand-truth when the base term of vocal is initialized by the overtone structure. Therefore it can be said that proposed method is better than the conventional method and NMF without constraint condition in both results of SDR and comparing the shapes of spectrogram.

V. Conclusion

We proposed a new NMF method based on harmonic, percussive and vocal constraint. The harmonic and vocal structures have a characteristic of temporal continuity, and the percussive structure has a characteristic of frequency continuity. In NMF, we added the harmonic and vocal constraint conditions to the activity term and percussive constraint condition to the base term. This is because the temporal information is owned by the activity term and the frequency information is owned by the bas term. And as post-processing, we removed low frequency from estimated vocal spectrogram.

As a result, we obtained better results than the conventional method for all of music pieces. The best result in SDR is due to proposed method by vowel sample base. The proposed method has better performance in recovering the vocal spectrogram when the overtone structure is used for the vocal's base spectra.

In the future, we may improve further by adapting cut-off frequency according to the characteristics of the input. Because the lowest frequency of vocal differs each music. And we hope to create method to perform these three steps simultaneously.

References

[1] S. Venkataramani, N. Nayak, P. Rao, and R. Velmurugan, "Vocal separation using singer-vowel priors obtained from polyphonic audio," pp. 283–288, 2014.

[2] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," pp. 90–93, Oct 2005.

[3] W.-H. Tsai and H.-M. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 330–341, Jan 2006.

[4] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 1475–1487, May 2007.

[5] Y.-G. Zhang and C.-S. Zhang, "Separation of music signals by harmonic structure modeling," *Advances in Neural Information Processing Systems*, pp. 1617–1624, 2006.

[6] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Signal Processing Conference, 2008 16th European*, pp. 1–4, Aug 2008.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," pp. 556–562, 2000.

[8] A. Chanrungutai and C. Ratanamahatana, "Singing voice separation for mono-channel music using non-negative matrix factorization," in *Advanced Technologies for Communications, 2008. ATC 2008. International Conference on*, pp. 243–246, Oct 2008.

[9] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Comput.*, vol. 19, pp. 780–791, Mar. 2007.

[10] C. Fevotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, March 2009.

[11] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," *Signal Processing Letters, IEEE*, vol. 21, pp. 1197–1200, Oct 2014.

[12] J. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, pp. 1180–1191, Oct 2011.

[13] L. Beranek, "Digital synthesis of speech and music," *Audio and Electroacoustics, IEEE Transactions on*, vol. 18, pp. 426–433, Dec 1970.

[14] "A study of speech recognition based on inner structure." http://www.geocities.jp/onsei2007/wav_data51/wav_data51.html.

[15] C.-L. Hsu and J.-S. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 310–319, Feb 2010.