Reverberation-robust underdetermined source separation with non-negative tensor double deconvolution

Naoki Murata*, Hirokazu Kameoka*†, Keisuke Kinoshita†, Shoko Araki†, Tomohiro Nakatani†,
Shoichi Koyama*and Hiroshi Saruwatari*
*Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
†NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation,
3-1 Morinosato Wakamiya, Atsugi-shi, Kanagawa 243-0198, Japan

***Abstract*—Source separation using an ad hoc microphone array can be useful for enhancing speech in such applications as teleconference systems without the need to prepare special devices. However, the positions of the sources (and the microphones when using an ad hoc microphone array) can change during recording, thus violating the commonly made assumption in many source separation algorithms that the mixing system is time-invariant. This paper proposes an extension of the multichannel nonnegative matrix factorization (NMF) approach to deal with the problem of underdetermined source separation in time-variant reverberant environments. The proposed method models the mixing system as a non-negative convolutive mixture based on the concept of a "semi-time-variant system" to handle the reverberation in a room as well allowing for relatively small changes in the source/microphone positions. It also models the power spectrogram of each sound source using the convolutive NMF model to consider the local dynamics of speech.**

I. Introduction

An ad hoc microphone array is a microphone array that uses built-in microphones in portable devices such as laptops and smartphones as spatially distributed sensors in an ad hoc fashion. It is particularly noteworthy in that it allows a flexible and convenient way to acquire audio data compared with traditional microphone arrays, which usually require special devices. Since a traditional microphone array is typically arranged so that the built-in microphones are located close to each other, the interchannel difference in arrival time (or phase difference) of each sound source provides an important clue in such applications as sound source localization and source separation. By contrast, since an ad hoc microphone array allows each microphone to be spatially distributed, the interchannel level difference of each sound source can also be a useful clue. Motivated by this, this paper proposes a source separation method specifically tailored for ad hoc microphone arrays that relies on the interchannel level difference of each sound source.

The problem of separating the signals of individual sound sources from the signals observed with a microphone array can be considered an inverse problem. In an underdetermined case where the sources outnumber the microphones, this inverse problem generally has an infinite number of solutions. Thus, some reasonable assumption is usually needed to limit the range of possible solutions. In recent years, multichannel extensions of non-negative matrix factorization (NMF) have attracted particular attention after being proposed as a powerful approach for underdetermined source separation [1]–[4]. Note that multichannel extensions of NMF for an overdetermined case have also been proposed with notable success [6], [7]. NMF was originally applied to monaural source separation where the magnitude (or power) spectrogram of a mixture signal, interpreted as a non-negative matrix, is factorized into the product of two non-negative matrices. This can be interpreted as approximating the observed spectra at each time frame as a linear sum of basis spectra scaled by time-varying amplitudes, and amounts to decomposing the observed spectrogram into the sum of low rank spectrograms. Multichannel NMF is an extension of NMF to a multichannel input, which assumes the power spectrogram of each underlying sound source to have a low rank structure.

While most of these approaches are formulated on the assumption that the mixing system is time-invariant, this assumption does not necessarily hold when the relative positions of sources and microphones are likely to change during recording. In such situations, methods based on the time-invariant mixing system do not work very satisfactorily.

This paper proposes an extension of the multichannel NMF approach to deal with the problem of underdetermined source separation in time-variant reverberant environments. The proposed method models the mixing system as a non-negative convolutive mixture based on the concept of the "semi-time-variant system" to take account of the reverberation in a room as well as to allow for relatively small changes in the source/microphone positions. The method also models the power spectrogram of each sound source using the convolutive NMF model [13] to take account of the local dynamics in the time-frequency components of speech.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

II. Proposed model

A. Convolutive mixture in time-frequency domain

We start by describing the mixing system with I sources and J microphones as a convolutive mixture in the time-frequency domain. When the transfer systems between the sources and the microphones are linear time-invariant and their impulse responses are sufficiently shorter than the window length of the short-time Fourier transform (STFT), the signals obtained by the microphones can be represented as an instantaneous mixture of source signals in the STFT domain, which is adopted in many conventional BSS methods including multichannel NMF [1]–[4]. However, when reverberation comes into play, the length of the room impulse responses can be longer than the STFT window. In such cases, the signals observed at the microphones can be well approximated by the convolutive mixture model [8], [9]

$$\boldsymbol{y}_{k,l} = \sum_i \sum_n \boldsymbol{a}_{i,k,n} s_{i,k,l-n}, \tag{1}$$

where i, k denote the source and frequency indices and l, n denote the time frame indices, respectively. $\boldsymbol{y}_{k,l} \in \mathbb{C}^J$ is a vector consisting of the complex time-frequency components observed at the J microphones. We hereafter use j to denote the microphone index. $\boldsymbol{a}_{i,k,n}$ represents the steering vector of each source, which corresponds to the contributions of the direct component (when $n = 0$) and the reverberant components arriving at the microphones with a time delay of n frames (when $n \neq 0$). $\boldsymbol{s}_{i,k,l}$ represents the complex time-frequency component of source i.

B. Nonnegative tensor double deconvolution

In a general scenario, the positions of the sources (and the microphones in an ad-hoc microphone array scenario) can change during recording. To handle such situations, we must consider a time-variant system. To achieve a source separation algorithm that is robust against changes in the acoustic environment, one possible approach involves treating a time-variant factor as a latent variable to be marginalized out so that the algorithm becomes less sensitive to that factor. [11] proposes a speech dereverberation method robust to speaker's movements that follows this idea. This idea is also adopted in a blind source separation method for an ad hoc microphone array [12], which is designed to be robust against the sampling rate mismatch of the microphones of the array. The proposed method uses a similar approach to develop a robust source separation algorithm that is designed to be less sensitive to the movements of the microphones and sources. Specifically, when the positions of sources and microphones are time-variant, i.e., the steering vector $\boldsymbol{a}_{i,k,n}$ depends on time l, the mixing process in (1) is described as a time-variant system

$$\boldsymbol{y}_{k,l} = \sum_i \sum_n \boldsymbol{a}_{i,k,n,l} s_{i,k,l-n}. \tag{2}$$

We assume that the time-frequency component of each source independently follows a complex Gaussian distribution, as in [15], i.e., $s_{i,k,l} \sim \mathcal{N}_{\mathbb{C}}(0, P_{i,k,l})$ where $P_{i,k,l}$ is the power spectrogram of source i. Thus, the time-frequency components observed at the J microphones follow

$$\boldsymbol{y}_{k,l} \sim \mathcal{N}_{\mathbb{C}}\Big(\boldsymbol{0}, \sum_i \sum_n P_{i,k,l-n} \boldsymbol{a}_{i,k,n,l} \boldsymbol{a}_{i,k,n,l}^{\mathsf{H}}\Big). \tag{3}$$

Here, we decompose the time-variant steering vector $\boldsymbol{a}_{i,k,n,l}$ into its magnitude and phase parts

$$\boldsymbol{a}_{i,k,n,l} = \begin{bmatrix} |a_{1,i,k,n,l}| & & 0 \\ & \ddots & \\ 0 & & |a_{J,i,k,n,l}| \end{bmatrix} \begin{bmatrix} e^{j\phi_{1,i,k,n,l}} \\ \vdots \\ e^{j\phi_{J,i,k,n,l}} \end{bmatrix}. \tag{4}$$

We assume that the magnitude part of the steering vector is less sensitive to the changes in the acoustic environment made for instance by a slight movement of the microphones/sources and assume a mixing system in which $|a_{J,i,k,n,l}|$ is time-invariant whereas $\phi_{j,i,k,n,l}$ is time-variant. We hereafter refer to this type of mixture process as *semi-time-variant system*. Thus, by assuming $|a_{j,i,k,n,l}| = A_{j,i,k,n}$, (4) can be rewritten as

$$\boldsymbol{a}_{i,k,n,l} = \underbrace{\begin{bmatrix} A_{1,i,k,n} & & 0 \\ & \ddots & \\ 0 & & A_{J,i,k,n} \end{bmatrix}}_{\boldsymbol{A}_{i,k,n}} \underbrace{\begin{bmatrix} e^{j\phi_{1,i,k,n,l}} \\ \vdots \\ e^{j\phi_{J,i,k,n,l}} \end{bmatrix}}_{\boldsymbol{\psi}_{i,k,n,l}}. \tag{5}$$

By substituting $\boldsymbol{A}_{i,k,n}\boldsymbol{\psi}_{i,k,n,l}$ for $\boldsymbol{a}_{i,k,n,l}$ in (3), we obtain the following distribution

$$\boldsymbol{y}_{k,l} \sim \mathcal{N}_{\mathbb{C}}\Big(\boldsymbol{0}, \sum_{i,n} P_{i,k,l-n} \boldsymbol{A}_{i,k,n} \boldsymbol{\psi}_{i,k,n,l} \boldsymbol{\psi}_{i,k,n,l}^{\mathsf{H}} \boldsymbol{A}_{i,k,n}^{\mathsf{H}}\Big). \tag{6}$$

While the phase part of the steering vector is one of the most useful clues to source separation for traditional arrays with spatially proximate microphones, the magnitude part of the steering vector, which can be seen as the interchannel level differences between the direct and reverberant components of each source, can also be a useful clue when it comes to an ad hoc array with spatially distributed microphones (imagine a situation where each microphone is located very close to a different speaker). Here, we instead consider the phase part of the steering vector to be a nuisance parameter since it can vary sensitively according to the movements of the microphones/sources. Hence, we treat $\phi_{j,i,k,n,l}$ as a latent variable to be marginalized out and make the following two assumptions:

- $\phi_{j,i,k,n,l}$ and $\phi_{j',i,k,n,l}$ ($j \neq j'$ or $l \neq l'$) are statistically independent.
- $\phi_{j,i,k,n,l}$ follows a uniform distribution in $[0, 2\pi)$.

By marginalizing out $\phi_{j,i,k,n,l}$, $\mathbb{E}[\boldsymbol{\psi}_{i,k,n,l}\boldsymbol{\psi}_{i,k,n,l}^{\mathsf{H}}]$ becomes an identity matrix. Thus, we obtain the following distribution

$$y_{j,k,l} \sim \mathcal{N}_{\mathbb{C}}\Big(0, \sum_i \sum_n P_{i,k,l-n} A_{j,i,k,n}^2\Big). \tag{7}$$

Until now, no assumptions have been made as regards $\boldsymbol{P}_i = (P_{i,k,l})_{K \times L}$. The conventional multichannel NMF [1]–[3] models the power spectrogram $P_{i,k,l}$ as a product of two nonnegative matrices. This can be interpreted as representing the

power spectrum of the source at each time frame as the non-negative combination of spectrum templates. Although this assumption may be relatively accurate for musical instrument sounds, it may be inadequate when it comes to speech since speech is characterized not only by the instantaneous spectra but also by the continuous transitions of these spectra over time. Therefore, we consider it reasonable to treat the concatenation of the spectra of multiple frames rather than a single-frame spectrum as a basic element constituting the entire spectrogram of speech. Motivated by this, we employ the convolutive NMF model introduced in [13], [14] to model the power spectrogram of each source. Namely, the power spectrogram of each source is modeled as a convolution over time of spectrogram templates and the corresponding temporal activation functions

$$P_{i,k,l} = \sum_{m} \sum_{\tau=0}^{T-1} W_{i,k,m,\tau} H_{i,m,l-\tau}, \tag{8}$$

where m and τ denote the indices of the spectrogram templates and time frame, respectively. $W_{i,k,m,\tau}$ and $H_{i,m,l}$ represent the ith spectrogram template and the corresponding temporal activation function. Note that this model reduces to the regular NMF model when $T = 1$.

Our proposed model involves two convolutions, one describing the reverberation process, and the other describing the power spectrogram model. Hence, the source separation problem based on the proposed model is formulated as a double deconvolution problem. We thus call our proposed method *Non-negative Tensor Double Deconvolution*, (NTDD).

III. Parameter estimation

From (7), the negative log-likelihood function of the unknown parameters is given by

$$\begin{aligned} C_{ML} &= \sum_{j,k,l} \log \mathcal{N}_{\mathbb{C}}\Big(y_{j,k,l}|0, \sum_{i,n} P_{i,k,l-n} A^2_{i,j,k,n}\Big) \\ &\stackrel{\mathrm{c}}{=} \sum_{j,k,l} d_{IS}\Big(|y_{j,k,l}|^2 | \sum_{i,n} P_{i,k,l-n} A^2_{i,j,k,n}\Big), \end{aligned} \tag{9}$$

where $\stackrel{\mathrm{c}}{=}$ denotes equality up to constant terms. $d_{IS}(y|x)$ denotes the Itakura-Saito divergence between x and y [15]. Therefore, the parameter estimation using the maximum likelihood estimation amounts to the minimization of the element-wise Itakura-Saito divergence between the obtained power spectrogram $Y_{j,k,l}$ and the modeled power spectrogram $X_{j,k,l} = \sum_{i,n} P_{i,k,l-n} A_{i,j,k,n}$ where $P_{i,k,l} = \sum_m \sum_\tau W_{i,k,m,\tau} H_{i,m,l-\tau}$.

Seen from a model-fitting perspective, it would be natural and interesting to consider employing divergence measures other than the Itakura-Saito divergence. Here, we generalize the divergence measure using the β divergence, which involves the Itakura-Saito divergence as a special case, and use the majorization-minimization (MM) approach to derive an iterative algorithm for finding the optimal parameters that minimize the β divergence between $Y_{j,k,l}$ and $X_{j,k,l}$.

We follow the same idea proposed in [17], [18] to construct an auxiliary function by bounding the convex term of the objective function from above using Jensen's inequality and bounding the concave term from above using a tangent line. Owing to space limitations, here we only derive an update rule for the magnitude part of the steering vector, namely $A_{j,i,k,n}$.

We can write the objective function as

$$\begin{aligned} J(\Theta) = &\frac{1}{\beta(\beta-1)} \sum_{j,k,l} Y^{\beta}_{j,k,l} + \frac{1}{\beta} \sum_{j,k,l} \Big(\sum_{i,n} A_{j,i,k,n} P_{i,k,l-n} \Big)^{\beta} \\ &- \frac{1}{\beta-1} \sum_{j,k,l} Y_{j,k,l} \Big(\sum_{i,n} A_{j,i,k,n} P_{i,k,l-n} \Big)^{\beta-1}, \end{aligned} \tag{10}$$

where Θ denotes the set of the parameters. Here, both the second and third terms of Eq. (10) involve the form $\frac{1}{\alpha}x^{\alpha}$ with $x = \sum_{i,n} A_{j,i,k,n} P_{i,k,l-n}$ and $\alpha = \beta, \beta - 1$. Note that $\frac{1}{\alpha}x^{\alpha}$ is concave in x on the interval $[0, \infty)$ when $\alpha \geq 1$ and convex when $\alpha < 1$. When $\frac{1}{\alpha}x^{\alpha}$ is convex, we can apply Jensen's inequality to build an upper bound function

$$\frac{1}{\alpha}x^{\alpha} \leq \frac{1}{\alpha} \sum_{i,n} \lambda_{j,k,l,i,n} \left(\frac{A_{j,i,k,n} P_{i,k,l-n}}{\lambda_{j,k,l,i,n}} \right)^{\alpha}, \tag{11}$$

where $\lambda_{j,k,l,i,n} \geq 0$ and $\sum_{i,n} \lambda_{j,k,l,i,n} = 1$. Equality holds when $\lambda_{j,k,l,i,n} = A_{j,i,k,n} P_{i,k,l-n} / \sum_{i,n'} A_{j,i,k,n'} P_{i,k,l-n'}$. Next, when $\frac{1}{\alpha}x^{\alpha}$ is concave, we can use the fact that a concave function lies below its tangent line to build an upper bound function

$$\frac{1}{\alpha}x^{\alpha} \leq Z^{\alpha-1}_{j,k,l} \Big(\sum_{i,n} A_{j,i,k,n} P_{i,k,l-n} - Z_{j,k,l} \Big) + \frac{Z^{\alpha}_{j,k,l}}{\alpha}, \tag{12}$$

where $x = Z_{j,k,l}$ denotes the point of tangency. Equality holds when $Z_{j,k,l} = \sum_{i,n} A_{j,i,k,n} P_{i,k,l-n}$. For simplicity of notation, we denote the right-hand sides of Eqs. (11) and (12) by $Q^{(\alpha)}_{j,k,l}(\Theta, \vartheta)$ and $R^{(\alpha)}_{j,k,l}(\Theta, \vartheta)$, respectively, where ϑ denotes the set of auxiliary variables $\{\lambda_{j,k,l,i,n}\}$ and $\{Z_{j,k,l}\}$. By using $R^{(\alpha)}_{j,k,l}(\Theta, \vartheta)$ and $Q^{(\alpha)}_{j,k,l}(\Theta, \vartheta)$, we arrive at the following auxiliary function

$$J^{+}(\Theta, \vartheta) = \frac{1}{\beta(\beta-1)} \sum_{j,k,l} Y^{\beta}_{j,k,l} + \sum_{j,k,l} S^{(\beta)}_{j,k,l}(\Theta, \vartheta), \tag{13}$$

where

$$S^{(\beta)}_{j,k,l} = \begin{cases} R^{(\beta)}_{j,k,l}(\Theta, \vartheta) - Y_{j,k,l} Q^{(\beta-1)}_{j,k,l}(\Theta, \vartheta) & \beta < 1 \\ Q^{(\beta)}_{j,k,l}(\Theta, \vartheta) - Y_{j,k,l} Q^{(\beta-1)}_{j,k,l}(\Theta, \vartheta) & 1 \leq \beta \leq 2 \\ Q^{(\beta)}_{j,k,l}(\Theta, \vartheta) - Y_{j,k,l} R^{(\beta-1)}_{j,k,l}(\Theta, \vartheta) & \beta > 2 \end{cases}. \tag{14}$$

According to the principle of the auxiliary function method, we can show that iteratively minimizing this function with respect to Θ and ϑ does not increase Eq. (10). The update equation for $A_{j,i,k,n}$ can thus be obtained as

$$A_{j,i,k,n} \leftarrow A_{j,i,k,n} \left(\frac{\sum_l Y_{j,k,l} X^{\beta-2}_{j,k,l} P_{i,k,l-n}}{\sum_l X^{\beta-1}_{j,k,l} P_{i,k,l-n}} \right)^{\varphi(\beta)}, \tag{15}$$

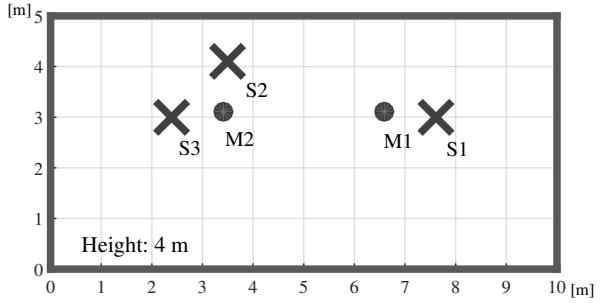

Fig. 1. Room configuration for numerical simulation

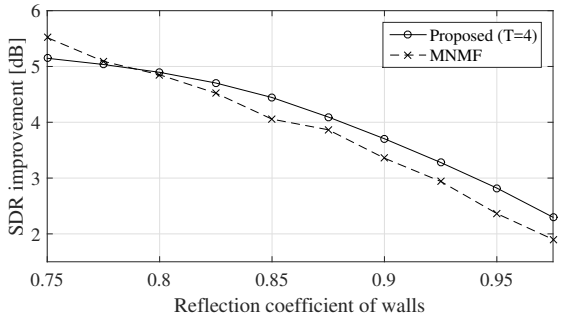

Fig. 2. SDR improvements of proposed method and MNMF with different settings of the reflection coefficient of walls

where $\varphi(\beta)$ is defined as

$$\varphi(\beta) = \begin{cases} 1/(2-\beta) & (\beta < 1) \\ 1 & (1 \leq \beta \leq 2) \\ 1/(\beta - 1) & (2 < \beta) \end{cases}. \quad (16)$$

Similarly, the update rules for other parameters are given as

$$W_{i,k,m,\tau} \leftarrow W_{i,k,m,\tau} \left(\frac{\sum_{j,l,n} Y_{j,k,l} X_{j,k,l}^{\beta-2} A_{j,i,k,l-n} H_{i,m,n-\tau}}{\sum_{j,l,n} X_{j,k,l}^{\beta-1} A_{j,i,k,l-n} H_{i,m,n-\tau}} \right)^{\varphi(\beta)},$$

$$H_{i,m,\tau} \leftarrow H_{i,m,\tau} \left(\frac{\sum_{j,k,l,n} Y_{j,k,l} X_{j,k,l}^{\beta-2} A_{j,i,k,l-n} W_{i,k,m,n-\tau}}{\sum_{j,k,l,n} X_{j,k,l}^{\beta-1} A_{j,i,k,l-n} W_{i,k,m,n-\tau}} \right)^{\varphi(\beta)}.$$

IV. Experiment

We conducted experiments on semi-blind source separation in reverberant environments to confirm the effectiveness of our proposed method. We used 45 utterances extracted from the ATR Japanese speech database [19] as the source signals. The number of speakers was 3, and the number of utterances per speaker was 15. First, to verify the robustness of the proposed method against reverberation, the proposed method was tested on synthetic signals simulating signals recorded in a room where the walls had different reflection coefficients. Second, to verify the robustness against the change in the positions of the microphones, the proposed method was tested on synthetic signals simulating signals recorded by moving microphones.

There were three sources and two microphones. The impulse responses were generated by using the image method. Figure 1 shows the two-dimensional configuration of the room. "•" (M1, M2) represent the microphone locations, and "×" (S1∼S3) represent the source locations. The signals of sources 1 and 2 were female utterances, and those of source 3 were male utterances. We compared the proposed method (Proposed) with the multichannel nonnegative matrix factorization (MNMF) proposed in [3]. Since MNMF is based on an anechoic (instantaneous) mixture model, its performance is expected to degrade as the reverberation time of the test data increases.

Out of the 15 utterances, one utterance was used for the separation and the remaining 14 utterances were used for pre-training the spectrogram templates of each source. We iterated the separation task 15 times for each selected utterance. For pre-training, convolutive NMF was used for the proposed method whereas NMF was used for MNMF. The frame number of each spectrogram template for both convolutive NMF and the proposed method was set experimentally at $T = 4$. The numbers of the templates of NMF and convolutive NMF were 40 and 20, respectively. The generalized KL divergence was adopted as a divergence measure for pre-training with NMF and convolutive NMF (where the parameter of the β divergence was $\beta = 1$). The STFT window length was 32 ms, and the shift length was 16 ms. We evaluated the source separation performance using the improvement of the signal-to-distortion ratio (SDR).

Figure 2 shows the SDR improvement results with the different reflection coefficients of the walls obtained with the proposed method and MNMF. The SDR improvements were averaged over the SDR improvements of each source on each microphone. For instance, the reverberation times (RT_{60}) were 620 ms, 720 ms and 1980 ms when the reflection coefficient of the walls were 0.75, 0.8 and 0.95, respectively. When the reflection coefficient of the walls was low, MNMF outperformed the proposed method. On the other hand, the proposed method outperformed MNMF as the reverberation time increased.

Figure 3 shows the SDR improvement results with different T settings. Note that when $T = 1$, the source spectrogram model reduces to the regular NMF model. As shown by the result, $T = 4$ outperformed $T = 1$, indicating that the convolutive NMF model is more suitable for expressing speech spectrograms than the regular NMF model.

Second, the proposed method was tested on synthetic signals that simulated signals recorded by moving microphones. The test signals were created by concatenating the signals recorded by the microphones located as shown in Figure 1 and by those moved a distance of Δx m along the x-axis. The reflection coefficient of the walls was set at 0.8. Figure 4 shows the

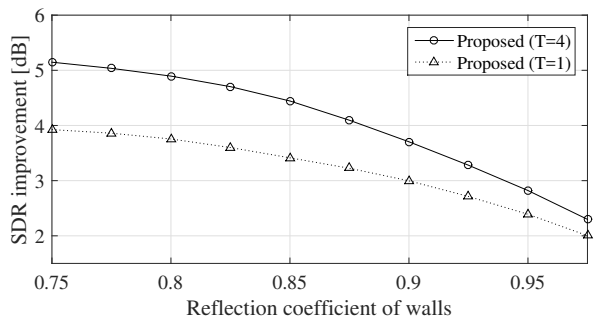

Fig. 3. SDR improvements of proposed method with different T settings

SDR improvement results with different Δx settings. The degradation of the MNMF performance, which assumes the time-invariance of the mixing system, was more significant than that of the proposed method as Δx increased. This may imply that the phase part of the steering vector is more sensitive to the microphone movements than the magnitude part.

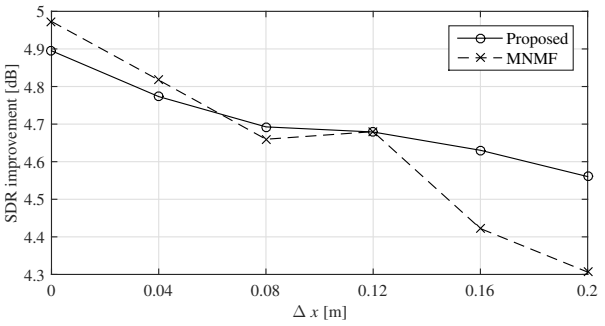

Fig. 4. SDR improvements of proposed method and MNMF with different Δx settings

V. Conclusion

This paper proposed an extension of the multichannel non-negative matrix factorization (NMF) approach to deal with the problem of underdetermined source separation in time-variant reverberant environments tailored for an ad hoc microphone array. The proposed method uses a non-negative convolutive mixture model to take account of the reverberation in a room as well as to allow for relatively small changes of the microphone/source positions. The method also uses the convolutive NMF model to express the power spectrogram of speech. We verified through experiments that the proposed method is robust against reverberation and microphone movements.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 26730100 and 26280060.

References

[1] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE TASLP, 18(3), 550–563, 2010.
[2] A. Ozerov et al., "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in Proc. ICASSP, 257–260, 2011.
[3] H. Sawada et al., "Multichannel extensions of non-negative matrix factorization with complex-valued data," IEEE TASLP, 21(5), 971–982, 2013.
[4] T. Higuchi and H. Kameoka, "Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model," in Proc. MLSP, 1–6, 2014.
[5] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in Proc. WASPAA, 177–180, 2003.
[6] H. Kameoka et al., "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in Proc. LVA/ICA, 245–253, 2010.
[7] D. Kitamura et al., "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in Proc. ICASSP, 276–280, 2015.
[8] T. Nakatani et al., "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in Proc. ICASSP, 85–88, 2008.
[9] T. Yoshioka et al., "Blind separation and dereverberation of speech mixtures by joint optimization," IEEE TASLP, 19(1), 69–84, 2011.
[10] S. Miyabe et al., "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain," in Proc. ICASSP, 674–678, 2013.
[11] H. Kameoka et al., "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in Proc. ICASSP, 45–48, 2009.
[12] H. Chiba et al., "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in Proc. IWAENC, 203–207, 2014.
[13] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in Proc. ICA, 494–499, 2004.
[14] P. D. O'Grady and B. A. Perlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in Proc. MLSP, 427–432, 2006.
[15] C. Févotte, N. Bertin, and J. L. Durrieu, "Non-negative matrix factorization with the Itakura-Saito divergence: With application to music analysis," Neural Computation, 21(3), 793–830, 2009.
[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Adv. NIPS, 556–562, 2001.
[17] H. Kameoka et al., "Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelopes," in IPSJ SIG Technical Reports, 2006-MUS-66-13,77–84, in Japanese, Aug. 2006.
[18] M. Nakano et al., "Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence," in Proc. MLSP, 283–288, 2010.
[19] A. Kurematsu, et al., "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, 9(4), 357–363, 1990.