On microphone arrangement for multichannel speech enhancement based on nonnegative matrix factorization in time-channel domain

Yoshikazu Murase*, Hironobu Chiba*, Nobutaka Ono†‡, Shigeki Miyabe*, Takeshi Yamada*, and Shoji Makino*
* University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
E-mail: {murase, chiba}@mmlab.cs.tsukuba.ac.jp, {miyabe, maki}@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp
† National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430 Japan
‡ The Graduate University for Advanced Studies (Sokendai)
E-mail: onono@nii.ac.jp

***Abstract*—In this paper, we investigate the relationship between the way microphones are arranged and the degree to which speech is enhanced using the transfer-function-gain nonnegative matrix factorization (NMF), which is an amplitude-based speech enhancement method that is suitable for use with an asynchronous distributed microphone array. In an asynchronous distributed microphone array, recording devices can be placed freely and the number of devices can be easily increased. Therefore, it is important that to determine the optimum microphone arrangement and the degree to which the performance is improved by using many microphones. We understood experimental evaluations to show that the performance by supervised NMF can achieve close to the ideal time-frequency masking with a sufficient number of microphones. We also show that the performance is better when more microphones are placed close to each source.**

I. Introduction

Recently, increasing attention has been paid to the asynchronous microphone array, which is a new framework with which to expand the applicability of microphone array signal processing [1]–[7]. This framework treats the simultaneous recording of a sound scene by multiple independent recording devices as a multichannel observation for array signal processing. This asynchronous recording approach has various advantages. For example, an arbitrary recording device can be used and it is easy to construct a many-channel microphone array with commonly available portable recording devices such as smartphones, voice recorders and laptop computers. In addition, the number and placement of microphones are flexible and the arrangement can be optimized for high-quality recording with high signal-to-noise ratios (SNRs). However, with conventional speech enhancement using asynchronous distributed microphone arrays, the speech enhancement performance deteriorates because of the phase difference caused by the sampling frequency mismatch between recording devices [1], [2]. One straightforward approach is to compensate the synchronization [3], [4]. However, the performance of speech enhancement is strongly affected by the synchronization accuracy. Another approach is to employ speech enhancement in the amplitude-spectrum domain discarding the phase for the sake of the robustness against synchronization error. Kako *et al.* proposed amplitude-spectrum beamformer for an asynchronous distributed microphone array, utilizing the target and non-target power ratio [5]. Togami *et al.* proposed a method that uses nonnegative matrix factorization (NMF) to estimate the transfer function gain (hereafter referred to as transfer-function-gain NMF) [8], [9], [10]. An amplitude-spectrum beamformer needs every single-source section containing the voice activity of only one speaker to learn the filter. In contrast, NMF has high potential for the unsupervised adaptation. Therefore, we focus on the transfer-function-gain NMF in this paper.

The goal of this study is to improve the speech enhancement performance of the transfer-function-gain NMF. The performance of transfer-function-gain NMF is affected by the number and position of microphones. Thus, in this paper we investigate what microphone arrangement is better or how much the performance is improved with many microphones. Experimental study reveals that the performance improves monotonically according to increase of microphones, but the improvement saturates at certain numbers of microphones and the benefit is limited. It is also revealed that placements with high SNRs are effective for speech enhancement.

II. Time-frequency masking with transfer-function-gain NMF

A. Problem statement

Preceding to the discussion of asynchronous observation, let us start with signal modeling of synchronized observation assuming that K sources are recorded by M microphones:

$$\begin{aligned}\mathbf{X}(\omega) &= [X_{mn}(\omega)]_{mn} \in \mathbb{C}^{M\times N}, \\ &= \mathbf{A}(\omega)\mathbf{S}(\omega), &(1)\\ \mathbf{A}(\omega) &= [A_{mk}(\omega)]_{mk} \in \mathbb{C}^{M\times K}, &(2)\\ \mathbf{S}(\omega) &= [S_{kn}(\omega)]_{kn} \in \mathbb{C}^{K\times N}, &(3)\end{aligned}$$

where, $[x_{ij}]_{ij} \in \mathbb{C}^{I\times J}$ is a matrix with the size $I \times J$ composed of the complex value x_{ij} in the ij-th entry. ω and N represent the frequency index and the number of time frames, respectively. $X_{mn}(\omega)$ is the signal observed at the

978-616-361-823-8 © 2014 APSIPA

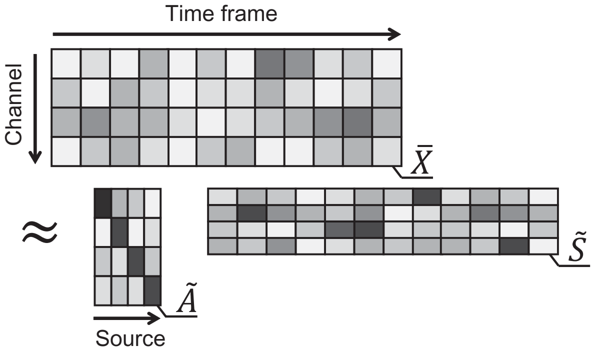

Fig. 1. Channel-time domain representation of observed signals for each frequency bin.

m-th microphone in the n-th time frame, $S_{kn}(\omega)$ is the k-th source signal in the n-th time frame, and $A_{mk}(\omega)$ is the transfer function from the k-th source to the m-th microphone.

We assume that the k-th microphone is placed at the closest position to the k-th source utilizing the flexibility of the asynchronous recording. In this case, the absolute value of $A_{kk}\ (k = 1, \dots, K)$ is at its highest in $A_{kj}, j = 1, \dots, K$.

Moreover, we assume that the phase of $X_{mn}(\omega)$ is not accurate due to the sampling frequency mismatch among devices, which causes phase drift. The purpose of this study is to enhance the signal with the highest SNR observed from each source and estimate the amplitude, $\bar{Y}_{kn} = |A_{kk}S_{kn}|$.

In the following all the modeling and processing can be carried out at each frequency bin. Therefore, we omit ω for simplicity.

B. Speech enhancement based on NMF

In this section, we describe the speech enhancement of the mixing model in the amplitude-spectrum domain, which uses NMF to estimate the parameter of the model. The parameterization of the NMF is shown in Fig. 1.

Assuming the additivity of the amplitude in the frequency domain, the mixing model can be expressed by the product sum of the amplitude spectrum omitting the phase;

$$\begin{aligned} \bar{\mathbf{X}} &= \left[\bar{X}_{mn}\right]_{mn} \in \mathbb{R}_+^{M\times N}, \\ &\approx \bar{\mathbf{A}}\bar{\mathbf{S}}, && (4) \\ \bar{\mathbf{A}} &= \left[\bar{A}_{mk}\right]_{mk} \in \mathbb{R}_+^{M\times K}, && (5) \\ \bar{\mathbf{S}} &= \left[\bar{S}_{kn}\right]_{kn} \in \mathbb{R}_+^{K\times N}, && (6) \end{aligned}$$

where $[x_{ij}]_{ij} \in \mathbb{R}_+^{I\times J}$ is a matrix with the size $I \times J$ composed of the nonnegative value x_{ij} in the ij-th entry. $\bar{X}_{mn}$, $\bar{A}_{mk}$ and $\bar{S}_{kn}$ represent the absolute amplitude of the observed signal, the transfer function and the source signal, respectively. Under this model, $\bar{\mathbf{A}}$ and $\bar{\mathbf{S}}$ are estimated by NMF to enhance the speech.

NMF approximates a nonnegative matrix as two low rank nonnegative matrices as follows.

$$\bar{\mathbf{X}} \approx \tilde{\mathbf{X}} = \tilde{\mathbf{A}}\tilde{\mathbf{S}}. \tag{7}$$

Here the distance measure between $\bar{\mathbf{X}}$ and $\tilde{\mathbf{A}}\tilde{\mathbf{S}}$ can be customized to the task by choosing from variety of functions which NMF can minimize. The low-rank approximation with the minimum distance restricts the solution of $\tilde{\mathbf{S}}$ to be sparse, resulting as the estimation of the source amplitude with $\mathbf{S}$ accompanied with the identification of the transfer function gain with $\tilde{\mathbf{A}}$. In other words, $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{S}}$ estimates $\bar{\mathbf{A}}$ and $\bar{\mathbf{S}}$.

Similarly to conventional source separation such as frequency-domain independent component analysis (ICA), transfer-function-gain NMF suffers from a permutation ambiguity. The problem causes a change in the sequential order of the separation signals in each frequency bin. To avoid this problem, assuming the absolute value of $A_{kk}\ (k = 1, \dots, K)$ is highest in $A_{kj}, j = 1, \dots, K$, we set the initial value of $\tilde{\mathbf{A}}$ as

$$\bar{A}_{mk} = \begin{cases} 1 & (m = k), \\ \alpha & (m \neq k), \end{cases} \tag{8}$$

where, α is an arbitrary positive real number that satisfies $\alpha < 1$.

The speech enhancement signal $\tilde{Y}_{kn}$, which enhanced k-th source, is obtained by the observed signal X_{kn} of the k-th microphones and Wiener filtering-based masking as

$$\tilde{Y}_{kn} = X_{kn} \frac{\left(\tilde{A}_{kk}\tilde{S}_{kn}\right)^2}{\sum_{i=1}^{K}\left(\tilde{A}_{ki}\tilde{S}_{in}\right)^2}. \tag{9}$$

This Wiener filtering mitigates the error in $\tilde{\mathbf{S}}$ caused by the model mismatch of the linear modeling in the amplitude spectrum domain.

Although the signal estimation by NMF discussed above is effective when $M >> N$, the performance deteriorates when K approaches to M. To solve this problem, Togami *et al.* proposed a method that introduced a penalty term into the activation $\tilde{\mathbf{S}}$ to make the matrix sparse [9]. Also, Chiba *et al.* proposed a method for improving the performance, by which involves the transfer function gain in a single source section containing the voice activity of only one speaker [10].

III. Effective speech enhancement conditions of transfer-function-gain NMF

As we discussed above, the nonnegative constraint in NMF is insufficient for the effective signal estimation, and requires additional conditions such as the availability of much more microphones than the sources. In this section we qualitatively discuss the effective condition for the transfer-function-gain NMF from the viewpoints of the number of the microphones and their placement. Note that we focus on the simple NMF and omit the discussion on the sparseness constraint introduced in [9].

A. Effect of number of microphones on speech enhancement

When $M = K$, the signal estimation is difficult with NMF, which has trivial solutions, for example:

$$\tilde{\mathbf{A}} = \mathbf{I},\ \ \tilde{\mathbf{S}} = \bar{\mathbf{X}}, \tag{10}$$

TABLE I
EXPERIMENTAL CONDITIONS.

Number of sources	4
Sampling frequency	16 kHz
Frame length	4096 samples
Frame shift	2048 samples
Signal length for evaluation	10 sec
Signal length for supervised NMF training	10 sec
Divergence	I-divergence
α (initialization parameter)	0.25
Number of NMF iterations	200
Reverberation time	0.3 sec

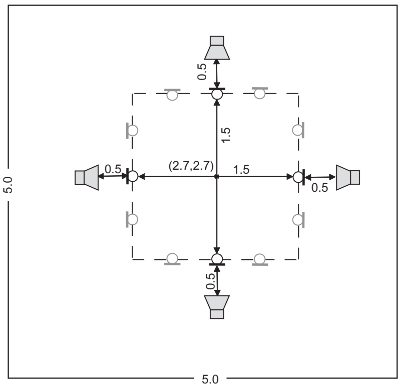

Fig. 2. Arrangement of microphones in experiment.

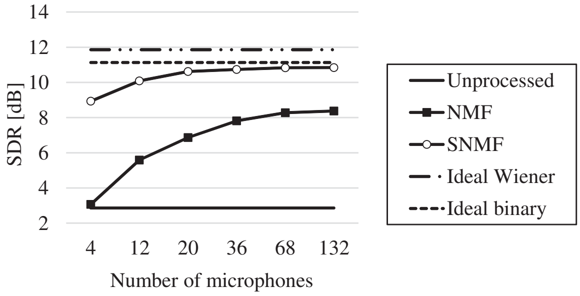

Fig. 3. SDRs of NMF and SNMF with increasing numbers of microphones.

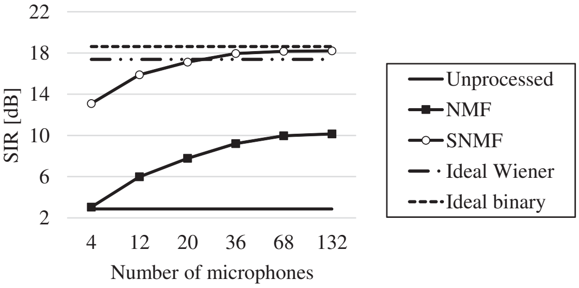

Fig. 4. SIRs of NMF and SNMF with increasing numbers of microphones.

where, $\mathbf{I} \in \mathbb{R}_+^{M \times M}$ is an identity matrix. The solutions shown in Eq. (10), satisfy $\tilde{\mathbf{X}} = \tilde{\mathbf{A}}\tilde{\mathbf{S}} = \bar{\mathbf{X}}$ and result in the error of the low rank approximation to be strictly zero. Thus, in this condition NMF usually converges to the trivial solution despite the insufficiency for the signal estimation. Therefore, when $M = K$, the solutions cannot enhance the speeches.

Similarly in the underdetermined condition with $M < K$, there exist trivial solutions satisfying $\tilde{\mathbf{X}} = \tilde{\mathbf{A}}\tilde{\mathbf{S}} = \bar{\mathbf{X}}$. Thus, NMF cannot conduct speech enhancement.

When $M > K$, the arbitrariness of the activation $\tilde{\mathbf{S}}$ is more regulated along with increase in the number of microphones. Therefore, the speech enhancement performance is improved monotonically because the estimated solutions differ from the trivial solutions.

B. *Effect of microphone arrangement on speech enhancement*

Another important factor of the transfer-function-gain NMF is microphone arrangement which varies the sparseness of the observation. When the observation is insufficiently sparse, NMF has arbitrariness of the solution and cannot provide the sufficient signal estimation as widely discussed in the field of multipitch analysis, e.g., in [11]. In the model of the transfer-function-gain NMF, the sparsity of the observation strongly depends on the microphone arrangement. If each one microphone observes one specific source with the high SNR, the observation maintains the high sparsity. Therefore, the effective microphone arrangement is in that each microphone observes one source intensively with the high SNR.

The SNRs affect both the unsupervised and supervised transfer-function-gain NMF. However, the effect on the supervised transfer-function-gain NMF is small because the basis $\tilde{\mathbf{A}}$ is estimated accurately in the pretraining process.

IV. EXPERIMENTAL EVALUATION OF SPEECH ENHANCEMENT PERFORMANCE WITH DIFFERENT NUMBERS OF MICROPHONES

A. *Experimental conditions*

Table I shows the experimental conditions. The observed signals were given as a convolutive mixture of clean speech and impulse responses with the image method [12]. To obtain the impulse response, we assumed all microphones to be non-directional. We conducted the experiment under six patterns where the microphones were placed uniformly on broken lines as shown in Fig. 2. Each pattern had 1, 3, 5, 9, 17 and 33 microphones at equal intervals on one broken line. One of the microphones is fixed in front of each speaker to record on a observed signal to enhance each source. Each pattern enhances the speech with totals 4, 12, 20, 36, 68 and 132 microphones.

Enhanced signals in the time domain are given by an inverse discrete-time-Fourier transform with the phase of the observed signal.

The signal-to-distortion ratio (SDR) and the source-to-interference ratio (SIR) are used as the evaluation scores [13]. The SDR evaluates the distortion of an enhancement signal, and the SIR evaluates the suppression ratio of non-target signals. The higher values of SDR and SIR show the better enhancement of the target source. We calculated the evaluation scores of 1) the unprocessed observation (Unprocessed), 2) the supervised transfer-function-gain NMF (SNMF), 3) the unsupervised transfer-function-gain NMF (NMF), 4) an Ideal

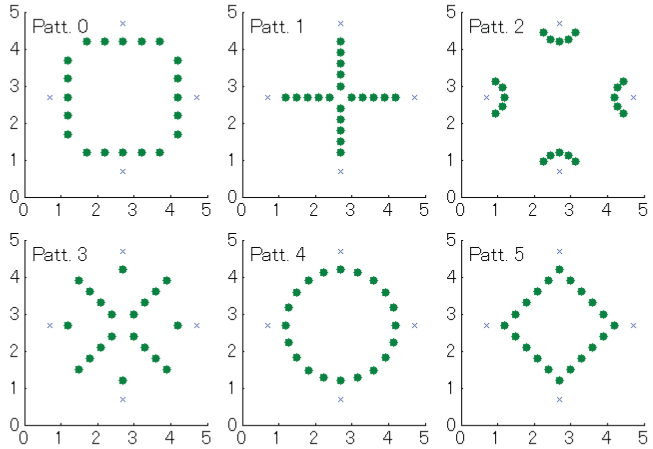

Fig. 5. Microphone arrangement of each pattern.

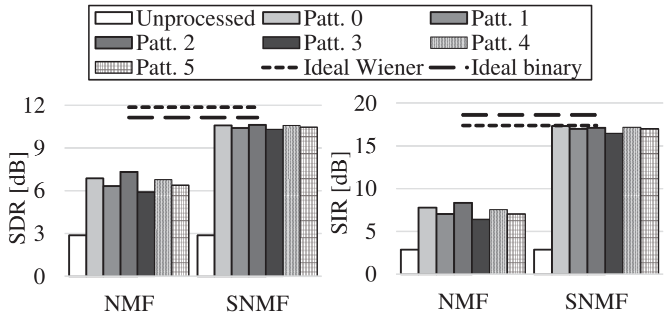

Fig. 6. SDRs and SIRs for each pattern with NMF and SNMF.

Wiener filter (Ideal Wiener) and 5) an ideal binary mask (Ideal binary).

B. *Results of evaluation experiment*

Figures 3 and 4 show the SDR and SIR values for each number of microphones. With NMF, the evaluation score with the 4 microphones is the same as that of Unprocessed. This result suggests that the estimated solutions converge on the trivial solutions. Thus, speech enhancement cannot be expected with NMF if the number of microphones is similar to the number of sound sources. With SNMF, the evaluation score with the 20 microphones approaches to those of the ideal binary mask, which shows the performance limit. The performance of both NMF and SNMF improves monotonically as the number of microphones increases, but the performance saturates. Therefore, even when the number of microphones is increased further, the performance of NMF cannot reach that of SNMF.

V. Experimental Evaluation of Speech Enhancement Performance with Various Arrangements of Microphones

A. *Experimental condition*

Figure 5 shows six microphone arrangement patterns. All patterns are constructed by twenty microphones. With all the patterns, four microphones are arranged in front of each speaker to observe with the high SNR. The sixteen remaining microphones are arranged for each pattern.

Pattern 0 corresponds to the arrangement with 20 microphones in section 3, pattern 1 arranges the microphones equally between two sound sources, and the arrangement does not overlap at the center, pattern 2 arranges the microphones in fan shapes and each sound source is at the center, pattern 3 arranges the microphones in diagonal lines and patterns 4 and 5 arrange the microphones in a circle and in a lozenge shape, respectively. In patterns 1 and 3, few microphones have single dominant sources, and the sparseness of the observed signals is hardly maintained. pattern 2 gives the high SNR of one specific source at each of the microphones and patterns 0, 4 and 5 give the great change in SNRs between adjacent microphones. SDR and SIR are independently calculated by Unprocessed, SNMF, NMF, Ideal Wiener and Ideal binary. The experimental conditions are the same as those in Sect. 4.1 except for the number and arrangement of the microphones.

B. *Results of evaluation experiment*

Figure 6 shows an evaluation of the performance of each pattern. For both SDR and SIR with NMF, the best performance was provided by pattern 2, followed in order by pattern 0, pattern 4 and pattern 5, pattern 1 and pattern 3. SNMF gave the same result. As a result, it is confirmed that the high performance is obtained by the arrangement with the high SNRs of each predominant source yielding the high.

VI. Conclusion

In this paper, we investigated the characteristics of speech enhancement performance with transfer-function-gain NMF as a result of the number and placement of the microphones when assuming the use of asynchronous distributed microphone arrays. First, we explained the modeling of transfer-function-gain NMF and the speech enhancement method, and discussed the desirable observation conditions. Second, we investigated the effect of the number and placement of microphones on the speech enhancement performance by employing simulation experiments with the image method. As a result, we found that the performance is improved monotonically as the number of microphones, however the improvement is limited. Moreover, an arrangement that maintain the high SNR of all the microphones can improve the performance.

VII. Acknowledgement

This work was supported by Grant-in-Aid for Scientific Research (B) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 25280069).

References

[1] E. Robledo-Arnuncio, T. S. Wada and B.-H.Juang, “On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation,” *Proc. WASPAA*, pp. 34–37, 2007.

[2] Z. Liu, “Sound source separation with distributed microphone arrays in the presence of clock synchronization errors,” *Proc. IWAENC*, 2008.

[3] S. Miyabe, N. Ono and S. Makino, “Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain, ” *Proc. ICASSP*, pp. 674-678, 2013.

[4] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada and S. Makino , “Speech enhancement with ad-hoc microphone array using single source activity, ” *Proc. APSIPA*, pp. 1-6, 2013.

[5] T. Kako, K. Kobayashi and H. Ohmuro, “A proposal of amplitude-spectrum beamformer for an asynchronous distributed microphone array, ” *Proc. Acoustic Society of Japan Spring Meeting*, pp. 829-830, Mar. 2013. (in Japanese)

[6] M. Souden, K. Kinoshita, M. Delcroix and T. Nakatani, “Distributed microphone array processing for speech source separation with classifier fusion,” *Proc. MLSP*, 2012.

[7] A. Bertrand, “Applications and trends in wireless acoustic sensor networks: a signal processing perspective,” *Proc. SCVT*, 2011.

[8] D. D. Lee and H. S. Seung, “Algorithms for nonnegative matrix factorization, ” *Proc. NIPS*, pp. 556-562, 2000.

[9] M. Togami, Y. Kawaguchi, H. Kokubo and Y. Obuchi, “Acoustic echo suppressor with multichannel semi-blind Non-negative matrix factorization, ” *Proc. APSIPA*, pp. 522-525, 2010.

[10] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada and S. Makino, “Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording, ” *Proc. IWAENC*, pp. 204-208, Sept. 2014.

[11] F. Rigaud, A. Falaize, B. David and L. Daudet, “Does inharmonicity improve an NMF-based piano transcription model?, ” *Proc. WASPAA*, 2013.

[12] E. A. P. Habets, “Room impulse response (RIR) generator,” Available: http://home.tiscali.nl/ehabets/rir_generator.html, Oct. 2008.

[13] E. Vincent, R. Gribonval and C. Fevotte, “Performance measurement in blind audio source separation, ” *IEEE Trans. on Audio, Speech & Language Processing*, vol.14, no. 4, pp. 1462-1469, 2006.