AMPLITUDE-BASED SPEECH ENHANCEMENT WITH NONNEGATIVE MATRIX FACTORIZATION FOR ASYNCHRONOUS DISTRIBUTED RECORDING

Hironobu Chiba[1], *Nobutaka Ono*[2,3], *Shigeki Miyabe*[1], *Yu Takahashi*[4], *Takeshi Yamada*[1], *Shoji Makino*[1]

[1]University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
[2]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430 Japan
[3]The Graduate University for Advanced Studies (Sokendai)
[4]YAMAHA Corp., Shizuoka, 438-0192, JAPAN
chiba@mmlab.cs.tsukuba.ac.jp, onono@nii.ac.jp,
{miyabe, maki}@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp, yu.takahashi@music.yamaha.com

ABSTRACT

In this paper, we investigate amplitude-based speech enhancement for asynchronous distributed recording. In an ad-hoc microphone array context, it is supposed that different asynchronous devices record speech. As a result, the phase information is unreliable due to sampling frequency mismatch. For speech enhancement based on the amplitude information instead of the phase information, supervised nonnegative matrix factorization (NMF) is introduced in the time-channel domain. The basis vectors, which represents the gain of the transfer function from a source to each microphone, are trained in advance by using single source observation. The experimental evaluations show that this approach is well robust against the sampling frequency mismatch.

Index Terms— Speech enhancement, ad-hoc microphone array, sampling frequency mismatch, nonnegative matrix factorization, time-frequency masking

1. INTRODUCTION

Asynchronous distributed microphone arrays provide a framework in which to use multiple independent recording devices for multichannel observation. Since the arrangement of the recording devices is flexible in this framework, it is easy to place the microphones near target speakers for recording with a high signal-to-noise ratio (SNR). However, the phase difference between asynchronous recording devices changes with time during digital signal processing due to the deviation of the sampling frequency of these devices. Since the phase difference is generally essential information for array signal processing, the synchronization must be corrected when we perform conventional microphone array signal processing [1, 2]. Therefore, researchers have studied the synchronization correction of asynchronous recording to allow them to apply the conventional microphone array signal processing to asynchronous recording [3, 4]. In this case, array signal processing performance is affected by the errors that arise during synchronization. In this study, we focus on the analysis of gain, which is not very sensitive to synchronization errors because amplitude information is not very sensitive to the synchronization error. If we can assume that each microphone is placed near one speaker, the gain ratios are regarded as being unique to the sources. To utilize such gain differences for speech enhancement, we employ nonnegative matrix factorization (NMF) where the transfer function gain is treated as the basis [5] (hereafter, transfer function gain NMF). By supervised the training of the basis using a single-source section, time-frequency masking with supervised NMF in the time-channel domain can realize robust speech enhancement of an asynchronous recording.

2. TIME-FREQUENCY MASKING WITH SUPERVISED NMF IN TIME-CHANNEL DOMAIN

2.1. Conventional linear mixing model

In this paper, a signal is expressed in the short-time Fourier transform (STFT) domain. Also, let $[x_{ij}]_{ij} \in \mathbb{C}^{I\times J}$ be a matrix with the size $I \times J$ such that the ij-th entry is a complex-valued x_{ij}.

In using a conventional synchronized microphone array, a multichannel observation can be expressed as

$$\mathbf{X}(\omega) = \mathbf{A}(\omega)\mathbf{S}(\omega), \quad (1)$$
$$\mathbf{X}(\omega) = [X_{mn}(\omega)]_{mn} \in \mathbb{C}^{M\times N}, \quad (2)$$
$$\mathbf{A}(\omega) = [A_{mk}(\omega)]_{mk} \in \mathbb{C}^{M\times K}, \quad (3)$$
$$\mathbf{S}(\omega) = [S_{kn}(\omega)]_{kn} \in \mathbb{C}^{K\times N}, \quad (4)$$

where ω is the frequency index, $X_{mn}(\omega)$ is the observed signal at the 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. Also, K, M and N represent the number of sources, the number of microphones, and the number of time frames, respectively.

Eq. (1) represents a conventional linear mixing model where $\mathbf{A}(\omega)$ represents a linear time-invariant mixing matrix.

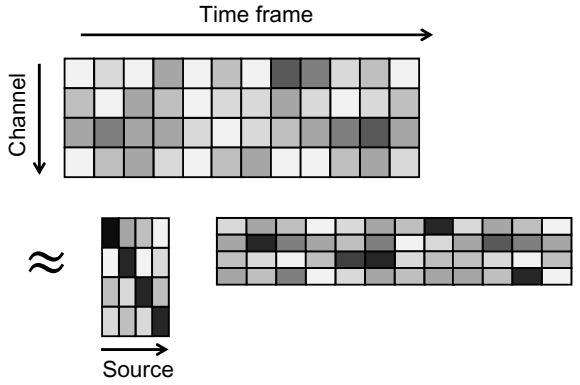

Fig. 1. Channel-time domain representation of observed signals for each frequency bin.

However, if we use an asynchronous distributed microphone array, Eq. (1) is not valid because the sampling frequency mismatch between devices causes phase drift, and then, $\mathbf{A}(\omega)$ can be time-varying. Therefore, we propose a model that works without the phase information, and which can even be applied to asynchronous recording.

2.2. Mixing model in amplitude domain

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

We assume an asynchronous recording case. Since the sampling frequency mismatch affects the phase, the amplitude of $\mathbf{A}$ (transfer function gain) can be time-invariant. Then, we assume that the linear mixing model in the amplitude domain in the following is approximately satisfied.

$$\bar{\mathbf{X}} \approx \bar{\mathbf{A}}\bar{\mathbf{S}} \tag{5}$$
$$\bar{\mathbf{X}} = \left[\bar{X}_{mn}\right]_{mn} \in \mathbb{R}_+^{M\times N} \tag{6}$$
$$\bar{\mathbf{A}} = \left[\bar{A}_{mk}\right]_{mk} \in \mathbb{R}_+^{M\times K} \tag{7}$$
$$\bar{\mathbf{S}} = \left[\bar{S}_{kn}\right]_{kn} \in \mathbb{R}_+^{K\times N} \tag{8}$$

Here, $\bar{X}_{mn}$, $\bar{A}_{mk}$ and $\bar{S}_{kn}$ represent the amplitudes of X_{mn}, A_{mk} and S_{kn}, respectively, and $\mathbb{R}_+$ represents a set of nonnegative real numbers. Note that such a mixing model in the power or the amplitude domain has frequently been assumed in the NMF context. So, we can also consider employing NMF for $\bar{\mathbf{X}}$ to estimate $\bar{\mathbf{S}}$ shown in Fig. 1. However, unlike NMF in the time-frequency domain, the number of microphones M does not greatly exceed the number of sources K in a usual case, and then, unsupervised NMF does not work well. Hence, we consider a supervised NMF [6] approach.

2.3. Supervised training of transfer function gain

To obtain the transfer function gain $\bar{\mathbf{A}}$ consisting of $\bar{A}_{mk}$, this paper employs the multiplicative update rule according to NMF with β-divergence [7, 8] .

$$\bar{A}_{mk} \leftarrow \bar{A}_{mk}\left(\frac{\sum_n \bar{X}_{mn}(\sum_k \bar{A}_{mk}\bar{S}_{kn})^{\beta-2}\bar{S}_{kn}}{\sum_n(\sum_k \bar{A}_{mk}\bar{S}_{kn})^{\beta-1}\bar{S}_{kn}}\right)^{\psi(\beta)} \tag{9}$$

$$\psi(\beta) = \begin{cases} \frac{1}{2-\beta} & \beta < 1 \\ 1 & 1 \le \beta \le 2 \\ \frac{1}{\beta-1} & \beta > 2 \end{cases} \tag{10}$$

Incidentally, the update rule corresponds to the Itakura-Saito divergence, I divergence, and Frobenius norm, respectively, with $\beta = 0, 1, \text{and } 2$. Moreover, $\bar{A}_{mk}$ and $\bar{S}_{kn}$ are normalized in each update by

$$\bar{A}_{mk} \leftarrow \frac{\bar{A}_{mk}}{\sum_m \bar{A}_{mk}}, \tag{11}$$
$$\bar{S}_{kn} \leftarrow \left(\sum_m \bar{A}_{mk}\right)\bar{S}_{kn}. \tag{12}$$

Since the transfer function gain NMF uses the transfer function gain associated with each unique speaker as a basis, we can assume that a solution close to the optimum one is obtained by training the transfer function gain in a single-source section containing the voice activity of only one speaker. Therefore, we obtain the basis matrix of the transfer function gain, $\bar{\mathbf{a}}_k = [\bar{a}_{mk}]_{m1} \in \mathbb{R}_+^{M\times 1}$ with rank one, from the training in the single-source section of the k-th sound source with the update rule according to Eqs. (9) and (14), which is introduced in the next section. Then, by coupling these basis matrices of the transfer function gain associated with each sound source, we construct the whole basis matrix of transfer function gain $\bar{\mathbf{A}} = \left(\bar{\mathbf{a}}_1 \cdots \bar{\mathbf{a}}_K\right)$.

Incidentally, assuming that the k-th source is close to the k-th microphone, we set initial value of the basis matrix as

$$\bar{a}_{mk} = \begin{cases} 1 & (m = k) \\ \alpha & (m \ne k) \end{cases}, \tag{13}$$

where $\alpha < 1$ is the initialization parameter of the transfer function gain of nontarget signal and is an arbitrary positive real number that satisfies $\alpha < 1$.

2.4. Estimating source by supervised NMF

In estimating source activation $\bar{\mathbf{S}}$ by supervised NMF, transfer function gain $\bar{\mathbf{A}}$ obtained in advance is given as the initial basis matrix and only $\bar{\mathbf{S}}$ is obtained by applying the multiplicative update rule given by

$$\bar{S}_{kn} \leftarrow \bar{S}_{kn}\left(\frac{\sum_m \bar{X}_{mn}(\sum_k \bar{A}_{mk}\bar{S}_{kn})^{\beta-2}\bar{A}_{mk}}{\sum_m(\sum_k \bar{A}_{mk}\bar{S}_{kn})^{\beta-1}\bar{A}_{mk}}\right)^{\psi(\beta)}. \tag{14}$$

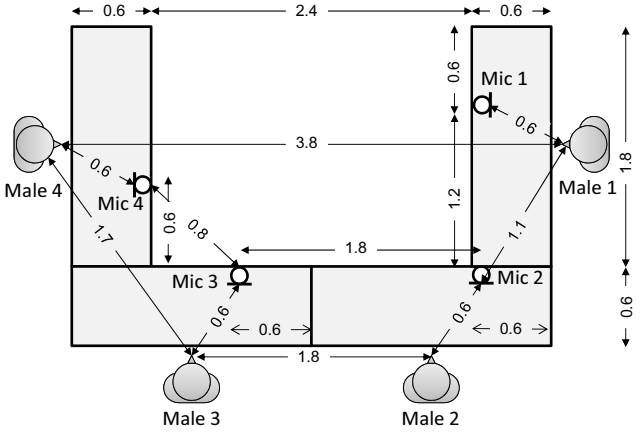

Fig. 2. Arrangement of microphones used in evaluation.

2.5. Time-frequency masking using Wiener filter

For speech enhancement in the time-frequency domain, the enhanced k-th target signal $\mathbf{Y}(k) = [Y_{mn}(k)]_{mn} \in \mathbb{C}^{M\times N}$ is obtained by

$$Y_{mn}(k) = W_{mn}(k)X_{mn}. \tag{15}$$

Here, the signal X_{mn} is obtained from the m-th microphone at the n-th time frame, and $W_{mn}(k)$ is the time-frequency domain mask that enhances the signal from the k-th sound source. In this study, the mask $W_{mn}(k)$ based on a Wiener filter in the time-frequency domain, and designed to enhance the target signal from the k-th sound source, is given according to the following equation using the estimated power of the signal from the k-th sound source, $\left(\bar{A}_{mk}\bar{S}_{kn}\right)^2$, as

$$W_{mn}(k) = \frac{\left(\bar{A}_{mk}\bar{S}_{kn}\right)^2}{\sum_k \left(\bar{A}_{mk}\bar{S}_{kn}\right)^2}. \tag{16}$$

3. EXPERIMENTAL EVALUATION

3.1. Experimental conditions

To confirm the performance of the time frequency masking with supervised NMF in the time-channel domain, we evaluated our proposed method using an actual conference speech recording given artificial sampling frequency mismatch. Both synchronous and asynchronous recorded data were used to demonstrate that our approach is robust to the phase difference caused by the difference in the sampling frequency between recording devices, as observed with asynchronous recording. The microphones and speakers were arranged as shown in Fig. 2. Table 1 shows a summary of the recording conditions used for the synchronous distributed microphone array. Here, we employed a relatively long frame since in general longer frame is more robust to sampling frequency mismatch. For quantitative evaluation, we first recorded speech of each single speaker with 16000 Hz sampling frequency, which was used as the ground truth. Then, synchronous recording data was generated by summing all of each source observation at each microphone. Asynchronous recording data were generated by resampling with sampling frequencies summarized in Table 2. The supervised training of NMF was performed by using a different part of single source observation from the evaluation. Table 3 shows a summary of the experimental conditions. The signal-to-distortion ratio (SDR) and source-to-interference ratio (SIR) were used as the evaluation measures [9]. SDR is a measure for evaluating the distortion of the output signal and SIR is a measure for evaluating the suppression of nontarget signals. As the SDR and SIR increase the target sound is further enhanced. SDR and SIR were calculated by the true source signals and the estimated source signals at each microphone. The SDR and SIR were independently calculated 1) without processing (Unprocessed), 2) by using the supervised transfer function gain NMF with the Itakura-Saito divergence (SNMF-ISdiv), 3) with the I divergence (SNMF-Idiv), and 4) with the Frobenius norm (SNMF-Fnorm), and 5) by independent vector analysis (IVA) [10]. The separation performance of IVA is considered to be low for asynchronous recording because IVA uses phase information to separate the sources.

Table 1. Recording environment

Microphones	SHURE SM57
Power amp.	YAMAHA XM4080
AD/DA conv.	Steinberg UR824

Table 2. Sampling frequencies of each microphone

Mic 1	16,000 Hz
Mic 2	16,001 Hz
Mic 3	16,002 Hz
Mic 4	16,003 Hz

Table 3. Experimental condition

Number of microphones	4
Number of sources	4
Frame length	4096 samples
Frame shift	2048 samples
Signal length for evaluation	10 sec
Signal length for supervised NMF training	10 sec
β-divergence	$0, 1, 2$
α (initialization parameter)	0.25
Number of NMF iterations	200

3.2. Results of evaluation experiment

Figs. 4 and 5 show the SDRs for synchronous and asynchronous recording, respectively. Figs. 6 and 7 show SIR values for synchronous and asynchronous recordings, respectively. For three types of supervised NMF, both the SDR and SIR were significantly higher than without processing. Moreover, the SDR and SIR for synchronous recording were equivalent to those for asynchronous recording, indicating that time-frequency masking using transfer function gain NMF is

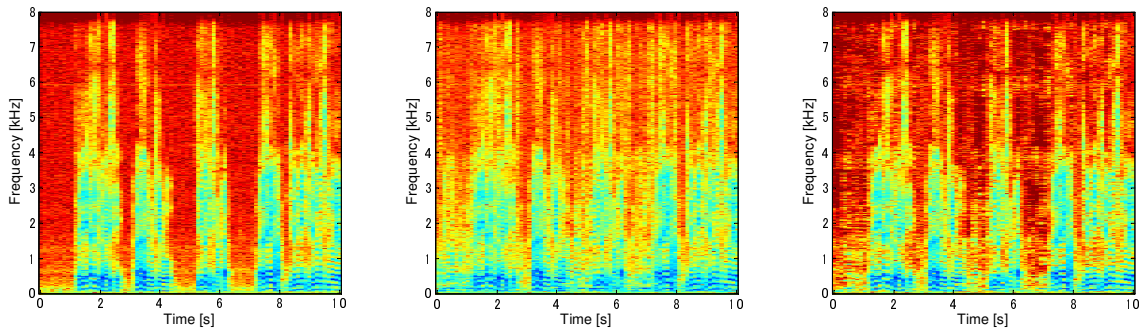

Fig. 3. Spectrograms of observed signal in mic 4: reference source (left), unprocessed (center), and enhanced signal with SNMF-Idiv (right).

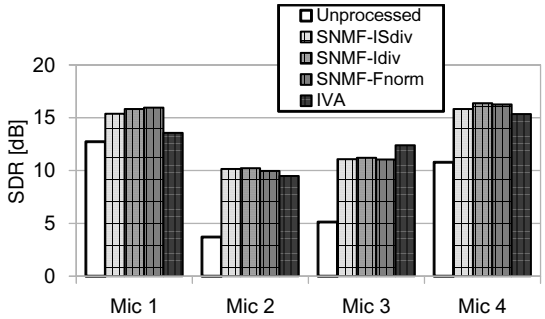

Fig. 4. The comparison of SDR between SNMFs and IVA in a synchronous recording case.

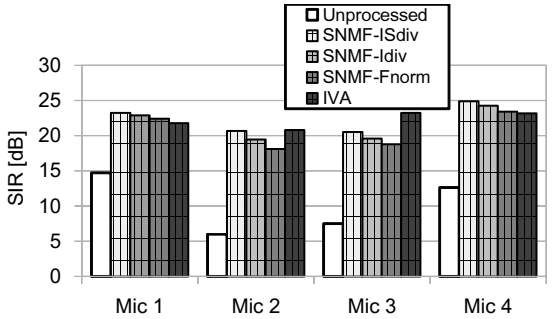

Fig. 6. The comparison of SIR between SNMFs and IVA in a synchronous recording case.

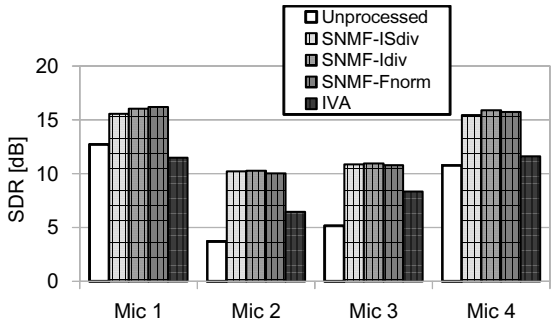

Fig. 5. The comparison of SDR between SNMFs and IVA in a asynchronous recording case.

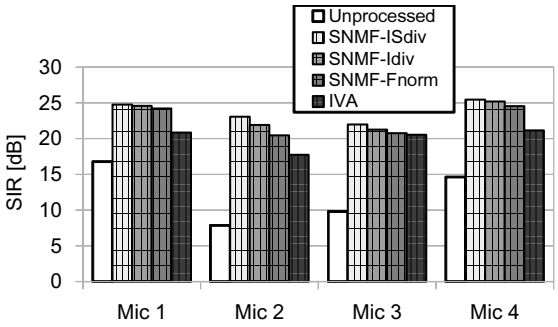

Fig. 7. The comparison of SIR between SNMFs and IVA in a asynchronous recording case.

robust to the phase differences between recording devices and appropriate for speech enhancement with an asynchronous distributed microphone array. The result of sound source separation by IVA indicated that the SDR and SIR for asynchronous recording were lower than those for synchronous recording. From this results, the separation performance of methods using phase information, as in IVA, was confirmed to be low for asynchronous recording.

4. CONCLUSION

The performance of time-frequency masking with supervised NMF in the time-channel domain was evaluated as a speech enhancement method that is robust to asynchronous recording. The results of experimental evaluations indicated that the SDR and SIR of the signals obtained by the proposed method were significantly improved compared with those of the observed signals. It was confirmed that this approach is robust to the phase difference and can be used as a speech enhancement method for asynchronous recording.

5. ACKNOWLEDGEMENT

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

6. 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] M. Togami, Y. Kawaguchi, H. Kokubo, and Y. Obuchi, "Sound source separation by utilizing amplitude-difference basis vectors of multiple sources," *Proc. Acoustical Society of Japan*, pp. 803–804, 2010.

[6] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. ICA*, pp. 414–421, 2007.

[7] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2007.

[8] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," *Proc. IEEE MLSP*, pp. 283–288, 2010.

[9] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[10] N. Ono, "Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique," *Proc. WASPAA*, pp. 189–192, 2011.