Multichannel NMF with Reduced Computational Complexity for Speech Recognition

Taiki Izumi*, Takanobu Uramoto*, Shingo Uenohara*, Ken'ichi Furuya*,
Ryo Aihara†, Toshiyuki Hanazawa† and Yohei Okato†,
* Faculty of engineering, Oita University
E-mail: v18e3001@oita-u.ac.jp
† Information Technology R&D Center, Mitsubishi Electric Corporation

***Abstract*—In this study, we propose efficient the number of computational iteration method of MNMF for speech recognition. The proposed method initializes and estimates the MNMF algorithm with respect to the estimated spatial correlation matrix reducing the number of iteration of update algorithm. This time, mask emphasis via Expectation Maximization algorithm is used for estimation of a spatial correlation matrix. As another method, we propose a computational complexity reduction method via decimating update of the spatial correlation matrixH. The experimental result indicates that our method reduced the computational complexity of MNMF. It shows that the performance of the conventional MNMF was maintained and the computational complexity could be reduced.**

I. Introduction

The use of voice-activated electronic devices has recently become widespread. However, voice recognition deteriorates in the presence of background noise, because sound, other than the target sound, enters the microphone. Research regarding sound source separation technology is underway to solve this problem.

In particular, nonnegative matrix factorization (NMF)[1] is a method that decomposes and analyses a matrix of nonnegative values. This can be applied to data such as sound, images, and sentences. In the field of acoustics, a multichannel extension has been proposed to consider spatial information of sound sources (MNMF)[2]. However, the computational complexity of MNMF increases with an increasing number of channels and it requires a long time to separate. In this study, we evaluate MNMF for speech recognition and propose a method for reducing computational complexity using an estimated spatial correlation matrix. In addition, we conducted a speech recognition experiment to demonstrate the effectiveness of this method.

II. MNMF Algorithm

MNMF decomposes an observation matrix $\mathbf{X}$ into four matrices ($\mathbf{H}$, $\mathbf{Z}$, $\mathbf{T}$, and $\mathbf{V}$) to realize source separation without prior learning. MNMF clusters spectral bases into L sources using spatial information[2].

A. Formuration

An observation vector was defined as $\tilde{\mathbf{x}} = [\tilde{x}_1, \cdots, \tilde{x}_M]^\top$, M the number of channels and $\top$ the transpose of the matrix.

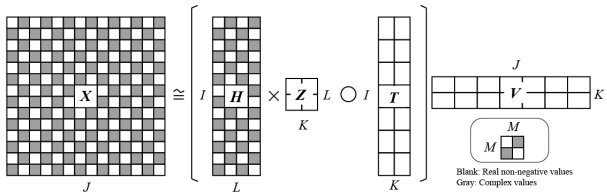

Fig. 1. Example of a decomposed matrix using MNMF (Gray denotes complex values)

Here, $\tilde{x}_m$ is the complex spectrum of the short-time Fourier transform at the mth microphone. At the frequency bin i ($1 \leq i \leq I$) and the time frame j ($1 \leq j \leq J$), an observation matrix $\mathbf{X}$ is represented as

$$\mathsf{X}_{ij} = \tilde{\mathbf{x}}_m \tilde{\mathbf{x}}_m^H = \begin{bmatrix} |\tilde{x}_1|^2 & \cdots & \tilde{x}_1 \tilde{x}_M^* \\ \vdots & \ddots & \vdots \\ \tilde{x}_M \tilde{x}_1^* & \cdots & |\tilde{x}_M|^2 \end{bmatrix} \tag{1}$$

where $*$ denotes the complex conjugate and ^{H} the Hermitian transpose. Matrix $\mathbf{X}$ is a hierarchical Hermitian positive semi-definite matrix whose elements are $M \times M$ complex matrices. Fig.1 shows that this matrix $\mathbf{X}$ is decomposed into four matrices. The basis matrix $\mathbf{T}(\in \mathbb{R}^{I \times K})$ consists of K bases, and the activation matrix $\mathbf{V}(\in \mathbb{R}^{K \times J})$ of the activation of each basis. The spatial correlation matrix $\mathbf{H}$ indicates the spatial information of sources of the sound, and the latent variable matrix $\mathbf{Z}(\in \mathbb{R}^{L \times K})$ associates the spatial information of the sources of the sound with each basis. Similar to $\mathbf{X}$, the matrix $\mathbf{H}$ is a hierarchical Hermitian positive semi-definite matrix whose elements are $M \times M$ complex matrices. This decomposition is defined as

$$\mathbf{X} \approx \hat{\mathbf{X}} = (\mathbf{HZ} \circ \mathbf{T})\mathbf{V} \tag{2}$$

where, $\circ$ denotes the Hadamard product. The right-hand side of the Eq.2 can be represented as

$$\hat{\mathbf{X}}_{ij} = \sum_{k=1}^{K} \left(\sum_{l=1}^{L} \mathsf{H}_{il} z_{lk} \right) t_{ik} v_{kj} \tag{3}$$

Ideally, $\hat{\mathbf{X}}$ whose elements are $\hat{\mathbf{X}}_{ij}$ matches with $\mathbf{X}$. However, in general, an error causes a discrepancy between them. To

calculate the difference between them, the Itakura–Saito (IS) divergence $D_{IS}(\mathsf{X}, \hat{\mathsf{X}})$ is employed as

$$D_{IS}(\mathsf{X}_{ij}, \hat{\mathsf{X}}_{ij}) = tr(\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}) - \log \det \mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1} - M$$

where $tr(\cdot)$ is the trace of a matrix.

III. Proposed Method

In this study, we propose a method to set up the number of computational iterations efficiently using spatial correlation matrix estimation and a method to decimate updates of the spatial correlation matrix, and subsequently reduced computational complexity.

A. Efficient iteration setting

In the first method, separation performance improves when using an initial value that estimates the spatial information in advance with regard to the spatial correlation matrix[3], [4]. Therefore, the conventional MNMF required approximately 500 updates to obtain the sufficient separation performance. However, performance is considered satisfactory with efficient iteration via setting the initial value to the spatial correlation matrix. Here, we improve the accuracy of the mask via setting the initial value of the mask emphasis using the Expectation Maximization algorithm (EM algorithm) by Nakatani *et al.* [5] as the mask of the target sound and noise generated by the binary mask. The steering vector (SV) of the target sound was estimated from the emphasized mask and the spatial correlation matrix was obtained and used as the initial value of MNMF.

B. Decimating update spatial correlation matrix

The second method estimates the spatial correlation matrix beforehand. It reducing computational complexity by decimating updates of the matrix **H**, which are expected to require the longest update times among the four matrices. The algorithm is shown in Fig.2.

C. Binary Mask

The binary mask[6] is a method of performing sound source separation by masking the time frequency based on the arrival time difference of each sound source. For example, the phase difference between the microphones is zero when the target sound source is directed forward. The phase difference becomes large when the noise arrives from zero degrees. The source of the target sound can be emphasized by masking the power of the time frequency bin where the phase difference between the microphones is away from zero. The mask M is set using a threshold value as follows:

$$W_{\omega,\ t} = \begin{cases} \epsilon & \text{if } |\theta_{\omega,\ t}| > \theta_c, \\ 1 & \text{if } |\theta_{\omega,\ t}| \leq \theta_c, \end{cases}$$

where, ω is a frequency bin, t the time frame, ϵ a sufficiently small constant, $\theta_{\omega,\ t}$ the phase difference of a time frequency bin, and θ_ϵ is a threshold value predetermined in advance.

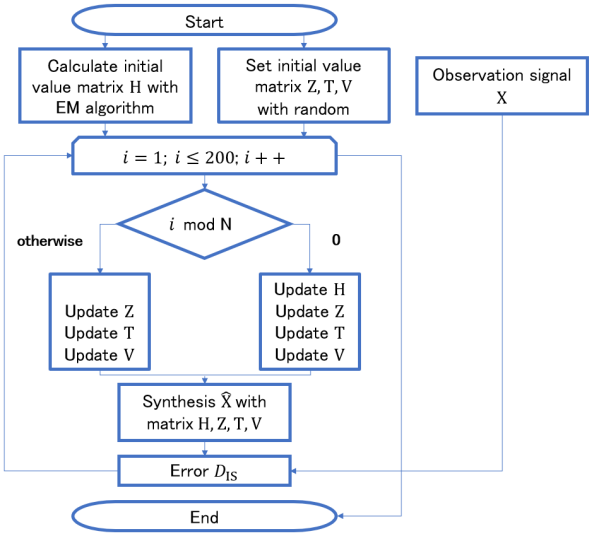

Fig. 2. Algorithm of reducing computational complexity

D. Mask Emphasis Based on EM Algorithm

Clustering of mask estimation is performed using complex GMM (CGMM)[7]. In CGMM, the observation signal vector $\tilde{\mathbf{x}}_{ij}$ has a complex Gaussian distribution. It is modelled as a mixture distribution

$$P(\tilde{\mathbf{x}}_{ij}; \theta) = \sum_n w_j^{(n)} N_c(\tilde{\mathbf{x}}_{ij}; 0, \sigma_{i,j}^{(n)} \mathbf{B}_j^{(n)}) \tag{4}$$

where n is an index that distinguishes the noise class $(n = v)$ from the speech + noise class $(n = x + v)$, $N_c(\tilde{\mathbf{x}}; \mu, \Sigma)$ is the mean μ, the complex Gaussian distribution of the covariance matrix Σ, and $w_j^{(n)}$ is the mixture ratio. It is assumed that the covariance dispersion matrix of each class can be decomposed into the product of the scalar value $\sigma_{i,j}^{(n)}$ and the matrix $\mathbf{B}_j^{(n)}$. where, θ represents a set of all model parameters. The maximum likelihood estimation is made the model parameters of CGMM based on the EM algorithm using the observation signal vector $\tilde{\mathbf{x}}_{ij}$ at all times for each frequency f. The following E-step and M-step were repeatedly applied to get an estimate.

1) E-step: Based on the estimated values of the model parameters of the CGMM obtained in M-step, the posterior probability that each time frequency point belongs to each class n is calculated as follows:

$$\hat{M}_{i,j}^{(n)} = \frac{\hat{w}_j^{(n)} N_c(\tilde{\mathbf{x}}_{ij}; 0, \hat{\sigma}_{i,j}^{(n)} \hat{\mathbf{B}}_j^{(n)})}{\sum_{n'} \hat{w}_j^{(n')} N_c(\tilde{\mathbf{x}}_{ij}; 0, \hat{\sigma}_{i,j}^{(n')} \hat{\mathbf{B}}_j^{(n')})} \tag{5}$$

the estimated value $\hat{M}_{i,j}^{(n)}$ of the mask is updated.

2) M-step: Based on the estimated value $\hat{M}_{i,j}^{(n)}$ of the mask obtained in the E-step, the estimated value of the model parameter of CGMM is updated as follows:

$$\hat{\mathbf{B}}_j^{(n)} = \frac{1}{\sum_t \hat{M}_{i,j}^{(n)}} \sum_t \hat{M}_{i,j} \frac{\tilde{\mathbf{x}}_{ij} \tilde{\mathbf{x}}_{ij}^H}{\hat{\sigma}_{i,j}^{(n)}} \tag{6}$$

$$\hat{\sigma}_{i,j}^{(n)} = \frac{1}{N}\tilde{\mathbf{x}}_{ij}^{H}(\hat{\mathbf{B}}_{j}^{(n)})^{-1}\tilde{\mathbf{x}}_{ij} \quad (7)$$

$$\hat{w}_{j}^{(n)} = \frac{\sum_t \hat{M}_{i,j}^{(n)}}{T} \quad (8)$$

where, N represents the number of microphones.
In the proposed method,we use the $M_{i,j}$ obtained by E-step.

E. SV Estimation Based on Binary Mask

Under the assumption that the target speech and noise are uncorrelated when the spatial correlation matrix $\mathbf{R}_j^{(x+v)}$ of the observation signal and the spatial correlation matrix $\mathbf{R}_j^{(v)}$ of the noise are known, the spatial correlation matrix $\mathbf{R}_j^{(x)}$ of the target speech can be obtained as follows:

$$\mathbf{R}_j^{(x)} = \mathbf{R}_j^{(x+v)} - \mathbf{R}_j^{(v)} \quad (9)$$

Also, each spatial correlation matrix using a mask can be obtained as follows:

$$\mathbf{R}_j^{(x+v)} = \frac{1}{J}\sum_{j=1}^{J}\tilde{\mathbf{x}}_{ij}\tilde{\mathbf{x}}_{ij}^{H} \quad (10)$$

$$\mathbf{R}_j^{(v)} = \frac{1}{\sum_{j=1}^{J} M_{t,v}^{(v)}}\sum_{j=1}^{J} M_{t,v}^{(v)}\tilde{\mathbf{x}}_{ij}\tilde{\mathbf{x}}_{ij}^{H} \quad (11)$$

where, $M_{t,v}^{(v)}$ is a mask indicating whether each time frequency point belongs to noise. SV can be approximated as the first eigenvector by obtaining the spatial correlation matrix $\mathbf{R}_j^{(x)}$ of the speech signal. Calculate the spatial correlation matrix from SV and use it as initial value of MNMF.

IV. Speech Recognition

A. Experimental condition

The effectiveness of the proposed method be confirming by speech recognition experiment. This time we used CHiME Challenge4[8] , which is a voice recognition task with speech recorded by 6 channel tablets equipped with 6 microphones in four noise environments (BUS, CAF, PED, STR). We used the word error rate (WER) to evaluate the performance of speech recognition. We used data recorded by six microphones. In addition, the target sounds of three types are prepared: learning set, development set, and evaluation set. There are real environment data (REAL) and virtual environment data (SIMU) in each. Here, we used a SIMU of development set. In the development set, data of 410 speech by four speakers are prepared in each environment. Parameters are shown in Table I. We compare the following methods.

1) Untreated (Noisy)
2) Delay sum array with weight (Baseline)
3) MNMF of 500 iterations with initial value random (500-Random)
4) MNMF of 500 iterations with initial value setting (500-EM)
5) MNMF of 200 iterations with initial value setting (200-EM)
6) MNMF of 200 iterations with initial value setting decimate the spatial correlation matrix at once every two times (200-EM-2)
7) MNMF of 200 iterations with initial value setting decimate the spatial correlation matrix at once every four times (200-EM-4)

TABLE I
Experimental conditions

Speech recognition system	Kaldi
Acoustic model	GMM
Vocabulary	5000
Language	English
Sampling frequency	16kHz
Frame size	1024
Shift size	256
Number of basis	30
Number of sound source	2

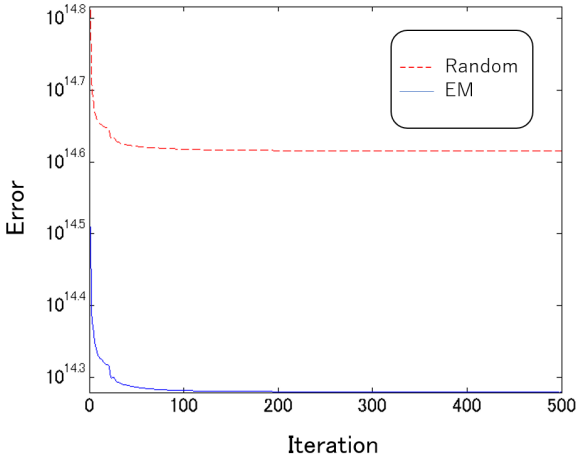

Fig. 3. Cost functions converges

B. Efficient Iteration

We considered setting an efficient number of iterations based on the convergence curve of the cost function when separating the 500 iterations. We observed that it converges with approximately 200 iterations from Fig.3. In addition, it is conceivable that it will result in a local optimal solution if the initial value is random. Therefore, we believe that a sufficient performance can be obtained with 200 iterations using the EM algorithm.

C. Computation time with increasing number of channels

Separation is conducted from 2-ch to 6-ch for a 14-s signal; the result is shown in Fig.4. We observed that the computation time increases exponentially with increasing number of channels (2 ch to 6 ch).

TABLE II
RESULT OF SPEECH RECOGNITION WER[%] (BOLDFACE IS THE BEST WER IN EACH ENVIRONMENTS)

-	BUS	CAF	PED	STR	AVE	Time
Noisy-1ch	20.38	29.81	20.49	27.30	24.49	-
Baseline-2ch	16.14	23.55	15.49	21.42	19.15	-
500-Random-2ch	23.0	32.1	25.2	29.9	27.5	301 s
500-EM-2ch	13.9	19.8	14.4	18.7	16.7	304 s
Baseline-6ch	12.74	17.29	11.80	15.56	14.35	-
500-Random-6ch	82.23	72.52	67.49	74.34	74.14	1966 s
500-EM-6ch	9.04	**11.81**	**8.55**	10.72	10.03	1971 s
(proposed 1) 200-EM-6ch	9.19	11.96	8.66	**10.32**	**10.03**	774 s
(proposed 2) 200-EM-2-6ch	**8.92**	12.91	9.00	10.86	10.42	640 s
(proposed 2) 200-EM-4-6ch	9.73	13.01	9.32	11.39	10.86	574 s

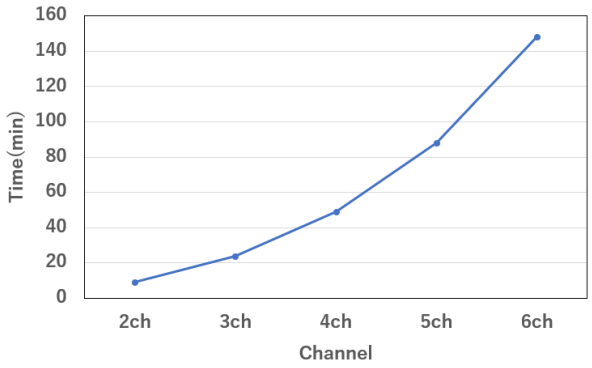

Fig. 4. Computation time with increasing number of channels

D. Recognition Experiment Result

Table II shows the results of MNMF with 2-ch and 6-ch as a comparison[9]. Time is the processing of MNMF for speech data of 6 seconds. Increasing the number of channels generally improves the WER based on the experimental results. The computation time was reduced to less than half compared with the 500 iterations case (500-EM-6ch), and the WER was not significantly affected because of reducing the number of iterations to 200. Regarding the thinning out method, the once in two updates WER was 0.39% worse but the computation time was 640 s. However, WER did not significantly deteriorate with respect to different environments. Similarly, the once in four updates WER was 0.83% worse but the computation time decreased to 574 s.

It shows that the performance of the conventional MNMF was maintained and the computational complexity could be reduced.

V. DISCUSSION

We confirmed that MNMF using the estimation regarding the spatial correlation matrix owing to EM algorithm is effective for improving speech recognition rate in noisy environments. It was observed that the method of decimating once every two times updates is effective only for a specific environment. It is believed that the difference between the results is that the difference between what should be optimized using the update formula of MNMF and what was estimated using EM algorithm depends on the environment.

VI. CONCLUSION

In this study, we examined the reducing computational complexity of MNMF using estimation of the spatial correlation matrix and confirmed its effectiveness through speech recognition experiments. It was confirmed that the estimation of the spatial correlation matrix using EM algorihtm is effective for improving the speech recognition rate in various noisy environments, and that speech recognition performance is not significantly affected even if fewer iterations are used for reducing computational complexity. It has been observed that reducing computational complexity using the method of decimating updates of the spatial correlation matrix is effective only in a specific environment. That is the performance of the conventional MNMF was maintained and the computational complexity could be reduced.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI grant numbers 16K00245 and 15H02728.

REFERENCES

[1] D.D. Lee, and H.S. Seung, “Learning the Parts of Objects with Non-negative Matrix Factorization,” Nature, vol.401, pp.788-791, 1999.
[2] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Multichannel Extensions of Non-Negative Matrix Factorization with Complex-Valued Data,” IEEE Trans. ASLP, vol.21, no.5, pp.971-982, 2013.
[3] I. Miura, *et al.*, “Analysis of Initial-value Dependency in Multichannel Nonnegative Matrix Factorization for Blind Source Separation and Speech Recognition” Jornal of IEICE, vol.J100-D, pp.376-384, 2017.
[4] T. Uramoto, *et al.*, “Sequential initialvalue setting associated with increasing number of channels in Multi-channel Nonnegative Matrix Factorization” ASJ, pp.535-538, 2017 Spring.
[5] T. Nakatani, *et al.*, “NTT CHiME-3 speech recognition system: noise-robust frontend” ASJ, pp.57-60,2016 Spring.
[6] H. Sawada, S. Araki and S. Makino, ”Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment”, IEEE Trans. Audio, Speech, Language Process., vol. 19 pp, pp. 516-527, Mar. 2011.
[7] N. Ito, S. Araki, T. Yoshioka and T. Nakatani, ”Relaxed disjointness based clustering for joint blind source separation and dereverberation”, in Proc. IWAENC, pp. 268-272, 2014.
[8] E. Vincent, S. Watanabe, J. Baker and R. Marxer, ”The 4th CHiME Speech Separation and Recognition Challenge”, CHiME CHALLENGE, http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016, Accessed May 2018.
[9] I. Miura, Y. Tachioka, T. Narita, S. Uenohara and K. Furuya, “Spatial Correlation Matrix Estimation Method in Initial Value Setting of Multi-channel Non-negative Matrix Factorization” ASJ, pp567-570,2017 Spring.