INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Coupled initialization of multi-channel non-negative matrix factorization based on spatial and spectral information

Yuuki Tachioka[1], Tomohiro Narita[1], Iori Miura[2], Takanobu Uramoto[2], Natsuki Monta[2], Shingo Uenohara[2], Ken'ichi Furuya[2], Shinji Watanabe[3], and Jonathan Le Roux[3]

[1]Information Technology R&D Center, Mitsubishi Electric Corporation, Japan
[2]Department of Computer Science and Intelligent Systems, Faculty of Eng., Oita University, Japan
[3]Mitsubishi Electric Research Laboratories, USA

Tachioka.Yuki@eb.MitsubishiElectric.co.jp

Abstract

Multi-channel non-negative matrix factorization (MNMF) is a multi-channel extension of NMF and often outperforms NMF because it can deal with spatial and spectral information simultaneously. On the other hand, MNMF has a larger number of parameters and its performance heavily depends on the initial values. MNMF factorizes an observation matrix into four matrices: spatial correlation, basis, cluster-indicator latent variables, and activation matrices. This paper proposes effective initialization methods for these matrices. First, the spatial correlation matrix, which shows the largest initial value dependencies, is initialized using the cross-spectrum method from enhanced speech by binary masking. Second, when the target is speech, constructing bases from phonemes existing in an utterance can improve the performance: this paper proposes a speech bases selection by using automatic speech recognition (ASR). Third, we also propose an initialization method for the cluster-indicator latent variables that couple the spatial and spectral information, which can achieve the simultaneous optimization of above two matrices. Experiments on a noisy ASR task show that the proposed initialization significantly improves the performance of MNMF by reducing the initial value dependencies.

Index Terms: non-negative matrix factorization, noisy speech, speech basis, spatial correlation, automatic speech recognition

1. Introduction

Source separation and noise reduction are essential techniques when processing real-world audio signals, whether the enhanced signals are to be listened by humans or further processed by machines. A prominent example is that of automatic speech recognition (ASR) in many hands-free applications. To increase the practicality of ASR systems, distant-talking input is much more desirable than close-talking input; however, noises or interferences significantly degrade the ASR performance.

One of the most effective methods is non-negative matrix factorization (NMF) [1, 2], which factorizes an observation matrix into two matrices: basis and activation matrices. To reconstruct the target signals from the mixed signals, it is important to properly construct the bases. Several methods have been proposed to construct proper initial bases for NMF: the k-means method [3], singular value decomposition [4, 5], and the LBG algorithm [6]. These methods only use training data for basis selection and it is possible that unnecessary bases are included because of a mismatch between training and test data. The small number of bases cannot represent speech well due to the large variety, because spectral properties of speech are dependent on speakers and utterances. The representation capability is improved by increasing the number of bases but it is then difficult to optimize them. Practically, it is necessary to restrict the number of bases and ideally to select bases that fit the phonemes appearing in the utterances in cooperation with ASR. One approach in this direction is ASR-assisted speech enhancement [7, 8, 9, 10], which seems to improve the performance. We extend the approach in [7] to a histogram-based one and validate the effectiveness on an ASR task.

Multi-channel NMF (MNMF) is a multi-channel extension of NMF, which is effective for source separation and noise reduction [11, 12], and factorizes an observation matrix into four matrices. It can consider both spatial and spectral information, simultaneously, by introducing Hermitian semi-positive definite matrices to handle phase information. The separation performance of MNMF is more dependent on initial values than NMF because the number of free parameters is larger.

The introduction of other methods or constraints helps to improve the performance of MNMF. The authors showed that the initial value dependencies are more dominant in the spatial correlation matrix than the other matrices and that its estimation using the cross-spectrum method is effective from enhanced speech by binary masking [13], whereas [14] showed the effectiveness of a rank-1 relaxation. Previous methods initialize bases and spatial correlation matrices, respectively, according to each criterion. However, these are coupled by cluster-indicator latent variables, thus, these spatial and spectral informations should be simultaneously exploited.

We propose effective initialization methods for MNMF parameters: ASR-based bases selection (Sec. 3.1), spatial correlation matrix initialization by using the cross-spectrum method and binary masking (Sec. 3.2), and combination of spatial and spectral information by cluster-indicator latent variables initialization (Sec. 3.3). This paper validates the effectiveness of the proposed method on the fourth CHiME challenge, a popular noisy ASR task [15], and analyzes the influence of each component in terms of the word error rate (WER).

2. Multi-channel non-negative matrix factorization (MNMF)

NMF factorizes an observation matrix $\boldsymbol{X}$ into two matrices: basis matrix $\boldsymbol{T}$ and activation matrix $\boldsymbol{V}$. In addition, MNMF factorizes an observation matrix $\boldsymbol{X}$ into four matrices $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$. The two additional matrices $\boldsymbol{H}$ and $\boldsymbol{Z}$ are the spatial correlation matrix and cluster-indicator latent variables, respectively. MNMF clusters K spectral bases into L sources by using the spatial information to achieve high source separation performance without any prior supervised training.

http://dx.doi.org/10.21437/Interspeech.2017-61

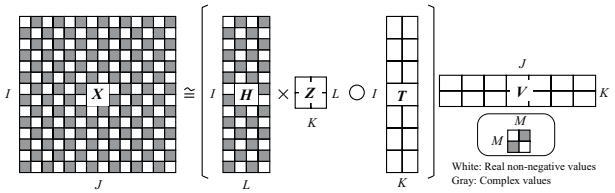

Figure 1: *An example of factorizing an observation matrix $\boldsymbol{X}$ into four matrices $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$ by the multi-channel NMF algorithm. ($I = J = 7$ and $K = L = M = 2$)*

2.1. Matrix factorization in MNMF

An observation vector $\boldsymbol{x}$ is $[x_1, \ldots, x_m, \ldots, x_M]^\top$ where $^\top$ denotes the transpose and x_m is the complex spectrum of the short-time Fourier transform (STFT), which is observed at the m-th microphone ($1 \leq m \leq M$). At the frequency bin i ($1 \leq i \leq I$) and the time frame j ($1 \leq j \leq J$), the element of an observation matrix $\boldsymbol{X} \in (\mathbb{C}^{M\times M})^{I\times J}$ is represented as

$$\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}(\boldsymbol{x}^*_{ij})^\top = \begin{bmatrix} |x_1|^2 & \cdots & x_1 x^*_M \\ \vdots & \ddots & \vdots \\ x_M x^*_1 & \cdots & |x_M|^2 \end{bmatrix}_{ij}, \quad (1)$$

where * denotes the complex conjugate. Matrix $\boldsymbol{X}$ is a hierarchical matrix whose elements $\boldsymbol{X}_{ij}$ are $M \times M$ complex Hermitian positive semi-definite matrices. MNMF factorizes this matrix $\boldsymbol{X}$ into four matrices $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$:

$$\boldsymbol{X} \cong \hat{\boldsymbol{X}} = [(\boldsymbol{H}\boldsymbol{Z}) \circ \boldsymbol{T}]\, \boldsymbol{V}, \quad (2)$$

where $\circ$ denotes the Hadamard product. Fig. 1 illustrates Eq. (2); $\boldsymbol{H} \in (\mathbb{C}^{M\times M})^{I\times L}$ is a spatial correlation matrix that indicates the spatial information of L sources and $\boldsymbol{Z} \in \mathbb{R}^{L\times K}$ is a cluster-indicator latent variables matrix that relates spatial information with each basis. Basis matrix $\boldsymbol{T} \in \mathbb{R}^{I\times K}$ is composed of K bases, and $\boldsymbol{V} \in \mathbb{R}^{K\times J}$ comprises the activations of each basis. The right-hand side of Eq. (2) can be represented as

$$\hat{\boldsymbol{X}}_{ij} = \sum_k \left[\sum_l \boldsymbol{H}_{il} z_{lk}\right] t_{ik} v_{kj}. \quad (3)$$

For ideal cases, the reconstructed matrix $\hat{\boldsymbol{X}}$ whose elements are $\hat{\boldsymbol{X}}_{ij}$ matches with the original matrix $\boldsymbol{X}$. However, in general, these matrices differ due to errors. In NMF, an arbitrary distance $D_*(\boldsymbol{X}, \hat{\boldsymbol{X}})$ between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ is defined and the above four matrices in the right-hand side of Eq. (2) are updated to minimize this distance. Here, the Itakura-Saito (IS) divergence[1]

$$d_{IS}(\boldsymbol{X}_{ij}, \hat{\boldsymbol{X}}_{ij}) = \mathrm{tr}(\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}^{-1}_{ij}) - \log\det \boldsymbol{X}_{ij}\hat{\boldsymbol{X}}^{-1}_{ij} - M, \quad (4)$$

is used, where $\mathrm{tr}(\cdot)$ is a trace of a matrix.

2.2. Multiplicative update rule

An iterative optimization algorithm, multiplicative update rule [17], is applied to the randomly initialized non-negative matrices $\boldsymbol{T}$, $\boldsymbol{V}$, and $\boldsymbol{Z}$, and the matrix $\boldsymbol{H}$ whose elements are initialized as unit matrices. These matrices are updated to minimize $D_{IS}(\boldsymbol{X}, \hat{\boldsymbol{X}})$ as follows:

$$\begin{aligned} t_{ik} &\leftarrow t_{ik}\sqrt{\frac{\sum_l z_{lk}\sum_j v_{kj}\,\mathrm{tr}(\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{H}_{il})}{\sum_l z_{lk}\sum_j v_{kj}\,\mathrm{tr}(\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{H}_{il})}}, \\ v_{kj} &\leftarrow v_{kj}\sqrt{\frac{\sum_l z_{lk}\sum_i t_{ik}\,\mathrm{tr}(\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{H}_{il})}{\sum_l z_{lk}\sum_i t_{ik}\,\mathrm{tr}(\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{H}_{il})}}, \\ z_{lk} &\leftarrow z_{lk}\sqrt{\frac{\sum_i\sum_j t_{ik}v_{kj}\,\mathrm{tr}(\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{H}_{il})}{\sum_i\sum_j t_{ik}v_{kj}\,\mathrm{tr}(\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{H}_{il})}}. \end{aligned} \quad (5)$$

$\boldsymbol{H}_{il}$ is a solution of an algebraic Riccati equation (6)

$$\boldsymbol{H}_{il}\mathsf{A}\boldsymbol{H}_{il} = \mathsf{B}, \quad (6)$$

whose coefficients A and B are

$$\begin{cases} \mathsf{A} = & \sum_k z_{lk}t_{ik}\sum_j v_{kj}\hat{\boldsymbol{X}}^{-1}_{ij}, \\ \mathsf{B} = & \boldsymbol{H}'_{il}\left[\sum_k z_{lk}t_{ik}\sum_j v_{kj}\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{X}_{ij}\boldsymbol{X}^{-1}_{ij}\right]\boldsymbol{H}'_{il}, \end{cases} \quad (7)$$

where $\boldsymbol{H}'_{il}$ is the value of matrix $\boldsymbol{H}_{il}$ before the update. The solution of Eq. (6) is found in the appendix of [12]. It is necessary to normalize matrices $\boldsymbol{H}$ and $\boldsymbol{Z}$, in order to preserve the uniqueness of Eq. (2) ($\boldsymbol{H}_{il} = \boldsymbol{H}_{il}/\,\mathrm{tr}(\boldsymbol{H}_{il})$) and the definition of probability ($z_{lk} = z_{lk}/\sum_l z_{lk}$).

Finally, the l-th separated source $\tilde{\boldsymbol{y}}_{ijl}$ ($1 \leq l \leq L$) can be obtained by the multi-channel Wiener filter as

$$\tilde{\boldsymbol{y}}_{ijl} = \left[\sum_k z_{lk}t_{ik}v_{kj}\right]\boldsymbol{H}_{il}\hat{\boldsymbol{X}}^{-1}_{ij}\boldsymbol{x}_{ij}. \quad (8)$$

3. Coupled initialization for MNMF

3.1. ASR-based initialization of speech bases $\boldsymbol{T}$

Fig. 2 shows the procedure of speech bases selection based on the ASR results. A total of K bases are composed of K_s speech bases and K_n noise bases. The noise bases are randomly initialized in the same manner as the conventional method.

Initial speech bases are sampled from the excerpt of the prepared clean speech. First, a basis dictionary is created from the clean speech data, where multiple frames are associated with each monophone. Monophone alignments are obtained by using ASR after converting the triphone alignments into monophone ones. The counts of each monophone are gathered in a histogram and the most frequent K_s monophones in an utterance are picked up from the dictionary. For each phoneme, each basis is selected randomly from the multiple frames in the dictionary.

In addition, some utterances that include more various phonemes need more bases than the other utterances. Then, it is possible to pick up the bases of frequently appearing monophones utterance-by-utterance by checking the appearance percentage, instead of selecting the fixed top K_s monophones. These two types of initializations are validated in the experimental section.

3.2. Initialization of spatial correlation matrices $\boldsymbol{H}$ using the cross-spectrum method

The separation performance can be improved by initializing $\boldsymbol{H}$ from impulse responses [13], but it is difficult to obtain these types of information a priori. The initial $\boldsymbol{H}$ can however be obtained from roughly separated sounds by using binary masking.

[1] IS divergence is suitable for the separation of music and speech, whose dynamic ranges are large [16].

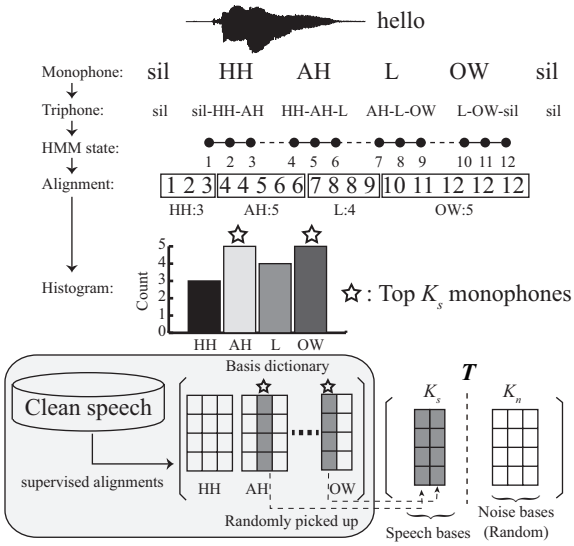

Figure 2: *Procedure of the ASR-based speech bases initialization ($K_s = 2$ and $K_n = 3$).*

3.2.1. Source signal enhancement by using binary masking

Binary masking is a source separation technique that masks spectra in the time-frequency domain based on the phase difference $\theta_{ij}(= \arg(x_2/x_1))$. For each source l, when a noise comes from another direction than that of the source, the phase difference will be different from that of the source, θ^s_{jl}. Each source can thus be enhanced by masking power spectra in the time-frequency bins that have different phases from θ^s_{jl}. The mask W_{ijl} can be set as

$$W_{ijl} = \begin{cases} \epsilon & (\min(|\theta_{ij} - \theta^s_{jl}|, 2\pi - |\theta_{ij} - \theta^s_{jl}|) > \theta_c), \\ 1 & (\text{otherwise}), \end{cases} \quad (9)$$

where $\epsilon(> 0)$ is a very small constant and θ_c is a threshold that can be set a priori. If the source direction is unknown, it can be estimated by various algorithms [18, 19].

3.2.2. Initialization by using the cross-spectrum method

The cross-spectrum method estimates the spatial correlation matrix at each frame, $\boldsymbol{H}_{ijl}$, as a multiplication of the l-th masked data and its Hermitian transpose [14]. After calculating $\boldsymbol{H}_{ijl}$, the initial $\boldsymbol{H}_{il}$ for MNMF is set as the expectation E_j of $\boldsymbol{H}_{ijl}$ in order for the estimations to be stable as shown in

$$\boldsymbol{H}_{il} = E_j\left[\boldsymbol{H}_{ijl}\right] = \frac{1}{\sum_j W^2_{ijl}} \sum_j W^2_{ijl} \boldsymbol{x}_{ij} \left(\boldsymbol{x}^*_{ij}\right)^\top. \quad (10)$$

3.3. Coupled initialization via cluster-indicator latent variables $\boldsymbol{Z}$

Cluster-indicator latent variables $\boldsymbol{Z}$ can explicitly relate the spatial information with the spectral information. Fig. 3 shows the system components of the proposed method. The combination of our methods described in Sections 3.1 and 3.2 provides the initial spatial correlation matrix $\boldsymbol{H}$ and the basis matrix $\boldsymbol{T}$. The left part of $\boldsymbol{H}$ is related to the target and its right part is related to the noise. In addition, the first K_s components of $\boldsymbol{T}$ are speech bases and the remaining ones are noise bases. To relate these matrices, the target parts of $\boldsymbol{Z}$ (the elements at the first row and the first to the K_sth columns) and noise parts of $\boldsymbol{Z}$ (the elements at the second row and the $(K_s + 1)$th to the Kth columns) should be set larger than the other parts of $\boldsymbol{Z}$. This initialization of $\boldsymbol{Z}$ strongly combines the target/noise spectral information derived from $\boldsymbol{T}$ and the target/noise spatial information derived from $\boldsymbol{H}$ to achieve their separation.

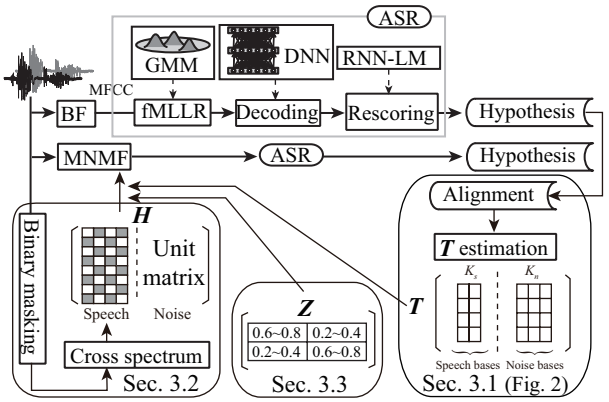

Figure 3: *Schematic diagram of the ASR system combined with the proposed initialization methods.*

4. Noisy ASR experiments (CHiME4)

4.1. Experimental setups

This paper validated the effectiveness of our proposed method on the 2ch track of the fourth CHiME challenge, a noisy ASR task with a vocabulary size of 5,000. The data were recorded by using hand-held tablets with six embedded microphones in four environments: bus (BUS), café (CAF), pedestrian (PED), and street (STR), with two types of data generation: data recorded in the real world (real) and data created by mixing real noise with clean speech recorded in a booth and convolved with measured impulse responses (simu). There are training, development (Dev), and test (Test) sets, and all the parameters for ASR were tuned on the Dev set.

We used the Kaldi toolkit [20]. The acoustic models were trained using the noisy data with no speech enhancement. The acoustic feature was the same as the challenge-provided one: the 13-dimensional MFCC $+\Delta + \Delta\Delta$ with feature-space maximum likelihood linear regression (fMLLR) transformation. These features were obtained by a first-pass decoding using Gaussian mixture model systems, and the features in 11 consecutive frames were concatenated and used as an input to the deep neural network (DNN). After the second-pass decoding using DNN systems, we used a recurrent neural network language model (RNN-LM) [21] for rescoring their hypotheses.

In the 2ch track, two channels were randomly sampled from the five channels with frontal direction[2]. Thus, microphone positions were different for every utterance. As conventional speech enhancement methods, we employed the challenge baseline beamformer (BeamformIt, denoted as BF) [22], as well as the minimum variance distortionless response (MVDR) beamformer with precise steering vector estimation [23]. The baseline was the conventional MNMF with random initialization of all matrices except $\boldsymbol{H}$, which was set to as a unit matrix [12]. There were two outputs of the conventional MNMF and it was necessary to select the appropriate one because it was unknown which one included the target speech. Here, this selection was oracle, i.e., the better hypotheses were selected according to the

[2]The total number of microphones was six but one microphone was located at the backend of the tablet.

Table 1: *Average WER [%] on the development and test sets of the fourth CHiME challenge for the baseline systems with conventional speech enhancement (SE) methods.*

SE	RNN-LM	Dev		Test	
		real	simu	real	simu
None	no	14.67	15.67	27.69	24.15
	yes	11.69	15.43	23.71	20.95
BF [22]	no	10.92	12.30	20.44	19.30
	yes	8.27	9.49	16.58	**15.39**
MVDR [23]	no	10.83	11.84	19.82	19.95
	yes	**7.91**	**9.35**	**15.91**	16.39

Table 2: *Average WER [%] for the proposed systems; MNMF denotes conventional MNMF where all initial matrices except for $\boldsymbol{H}$ which is set to a unit matrix are random. (I) uses a binary masking based $\boldsymbol{H}$ initialization. (II) is (I) with ASR-based speech bases selection. (III) is (II) with speech bases kept constant during the MNMF update. (IV) is (II) with $\boldsymbol{Z}$ initialization. (V) is (IV) with variable-size speech bases.*

SE	RNN-LM	Dev		Test	
		real	simu	real	simu
MNMF	no	23.99	22.42	33.98	23.18
	yes	20.96	19.62	23.46	19.49
(I)	no	10.54	10.83	18.80	15.93
	yes	7.84	8.32	14.83	12.72
(II)	no	10.16	10.68	18.63	16.87
	yes	7.53	**8.20**	14.98	13.75
(III)	no	11.00	11.16	19.65	16.03
	yes	8.08	8.68	15.91	12.42
(IV)	no	10.00	10.77	17.88	14.48
	yes	7.42	8.26	13.97	10.99
(V)	no	9.74	10.72	17.78	14.15
	yes	**7.30**	8.33	**13.81**	**10.91**

utterance-based WERs after both were decoded, which is the upper limit performance of the conventional MNMF. Parameter settings of MNMF were as follows: $I = 513$, $K = 30$, and $L = M = 2$, which was common through all the experiments.

For our $\boldsymbol{H}$ initialization, the binary masking assumed that the target speaker was in the frontal position and ideally, the phase differences of the target source θ^s were near zero, but some errors did occur. For our $\boldsymbol{T}$ initialization, K_s was set to be 20. For the selection involving the top candidates up to a given percentage, the percentage was set such that roughly 20 bases were used on average.

4.2. Results and discussions (Baseline and conventional methods)

Table 1 shows the baseline WERs of the challenge. Baseline BF significantly improved the performance over the unprocessed signals. RNN-LM rescoring reduced the errors by 20%. MVDR achieved equivalent performance with baseline BF, although in the 6ch track, MVDR outperformed BF [23, 24]. Table 2 shows the performance of the conventional MNMF with random initialization, which was even worse than those of the baselines due to spectral distortions introduced by the separation.

4.3. Results and discussions (Proposed methods)

Table 2 also shows the performance of the proposed method. $\boldsymbol{H}$ initialization ((I) in the table) significantly improved the performance, outperforming both BF and MVDR. Association with $\boldsymbol{T}$ initialization (II) further improved the WER by 0.2–0.4% on the Dev set. Keeping speech bases constant (III) did not improve the performance because there were mismatches between training and test data, thus, updating the bases is necessary. $\boldsymbol{Z}$ initialization (IV) gave additional improvements. Variable-size speech bases (V) improved the performance in some cases but this was not significant. Table 3 shows the WER of the respective methods per environment. Our approach was effective for all environments.

Table 3: *WER [%] per environment for each system with RNN-LM rescoring.*

SE	Envir.	Dev		Test	
		real	simu	real	simu
None	BUS	15.25	13.55	36.19	16.40
	CAF	12.18	19.46	24.58	24.09
	PED	7.51	11.11	19.77	20.53
	STR	11.81	17.62	14.33	22.79
BF	BUS	10.93	8.17	25.37	10.63
	CAF	8.14	12.11	15.89	18.27
	PED	5.19	7.17	13.60	15.67
	STR	8.82	10.58	11.45	16.83
(I)	BUS	9.59	6.92	22.81	8.20
	CAF	7.52	10.68	14.76	14.64
	PED	5.66	**6.34**	12.00	12.39
	STR	8.73	9.34	10.42	14.64
(IV)	BUS	9.78	7.29	21.95	7.71
	CAF	7.17	**10.43**	13.19	12.61
	PED	**5.10**	6.37	**10.30**	**11.21**
	STR	**7.61**	**8.97**	**9.86**	12.42
(V)	BUS	**8.91**	**6.90**	**21.28**	**7.55**
	CAF	**7.02**	10.96	**13.02**	**12.01**
	PED	5.35	6.50	10.91	11.23
	STR	7.90	9.16	10.03	**12.14**

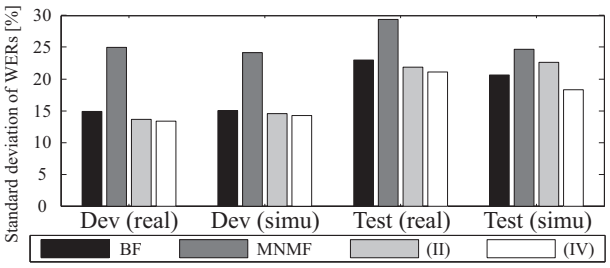

Figure 4: *Standard deviations of WERs for each method.*

Fig. 4 shows the standard deviations of the WERs for each speech enhancement. The conventional MNMF had significantly larger standard deviations than the others, which shows the large initial value dependencies. The proposed $\boldsymbol{T}$ initialization (II) decreased the standard deviations and combining $\boldsymbol{Z}$ initialization (IV) achieved the smallest standard deviation.

5. Conclusion

This paper proposed effective initialization methods for three of the four MNMF matrices. First, the basis matrices corresponding to the speech were initialized from the clean speech based on the ASR results. Second, the spatial correlation matrices were constructed from the sounds roughly separated by binary masking. Third, the cluster-indicator latent variables were initialized to combine the two matrices above. Experimental results on the fourth CHiME challenge show that these initializations were effective for noisy ASR. Compared with the baseline beamformer, although MNMF with random initialization did not improve the WERs, MNMF with the proposed initialization significantly improved the WER.

6. References

[1] D. D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[3] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithm*, 2007, pp. 1027–1035.

[4] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, pp. 1350–1362, 2008.

[5] H. Qiao, "New SVD based initialization strategy for non-negative matrix factorization," *Pattern Recognition Letters*, vol. 63, pp. 71–77, 2015.

[6] K. Kwon, J. W. Shiny, I. Choi, H. Y. Kim, and N. S. Kim, "Incremental approach to NMF basis estimation for audio source separation," in *Proceedings of APSIPA*, 2016, pp. 1–5.

[7] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent NMF for speech enhancement in monaural mixtures," in *Proceedings of INTERSPEECH*, 2011, pp. 1217–1220.

[8] F. Sohrab and H. Erdogan, "Recognize and separate approach for speech denoising using nonnegative matrix factorization," in *Proceedings of EUSIPCO*, 2015.

[9] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proceedings of INTERSPEECH*, 2015, pp. 1760–1764.

[10] W. ZQ and W. DL, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[11] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[12] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[13] I. Miura, Y. Tachioka, T. Narita, J. Ishii, F. Yoshiyama, S. Uenohara, and K. Furuya, "Analysis of initial-value dependency in multichannel nonnegative matrix factorization for blind source separation and speech recognition (in Japanese)," *IEICE Transactions on Information and Systems*, vol. J100-D, no. 3, pp. 376–384, 2017.

[14] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proceedings of ICASSP*, 2015, pp. 276–280.

[15] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, to appear, 2016.

[16] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation MIT Press*, vol. 21, pp. 793–830, 2009.

[17] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence," in *Proceedings of MLSP*, 2010, pp. 283–288.

[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[19] Y. Tachioka, T. Narita, and T. Iwasaki, "Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information," *Acoustical Science & Technology*, vol. 33, no. 1, pp. 68–71, 2012.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011, pp. 1–4.

[21] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of INTERSPEECH*, 2010, pp. 1045–1048.

[22] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2011–2023, 2007.

[23] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proceedings of ASRU*, 2015, pp. 436–443.

[24] Y. Tachioka, S. Watanabe, and T. Hori, "The MELCO/MERL system combination approach for the fourth CHiME challenge," in *Proceedings of the Fourth CHiME Challenge Workshop*, 2016, pp. 1–3.