Complex tensor factorization in modulation frequency domain for single-channel speech enhancement

Shogo Masaya[1,2], *Masashi Unoki*[1]

[1]Japan Advanced Institute of Science and Technology
[2]INPEX

masayaka9u5@gmail.com, unoki@jaist.ac.jp

10.21437/Interspeech.2015-410

Abstract

This paper proposes a novel method of speech enhancement using tensor factorization, which is extended from complex non-negative matrix factorization (CMF), in the modulation frequency domain. Non-negative matrix factorization (NMF) has attracted a great deal of attention as a recent approach to speech enhancement for its ease of feature detection in the acoustic frequency domain. However, previous studies have suggested that spectral processing like spectral subtraction in the modulation frequency domain has been an effective scheme for speech enhancement. The use of not only the amplitude information but also the phase information is required in the modulation frequency domain to utilize more information on speech. Thus, we present new tensor factorization on the complex spectrum in the modulation frequency domain for single-channel speech enhancement. The amplitude and phase spectrum in the acoustic frequency domain can be estimated by using the factorized complex spectra in the modulation frequency domain. Numerical experiments were carried out under several noisy conditions to evaluate the effectiveness of the proposed method. The signal to error ratio and signal to noise ratio loss were used as objective measures. The results revealed that the proposed method outperformed the existing methods of speech enhancement based on NMF and CMF.

Index Terms: Non-negative matrix factorization, complex non-negative matrix factorization, tensor factorization, modulation frequency domain, speech enhancement

1. Introduction

Speech enhancement techniques have potential applications in various fields such as in the front-ends of speech recognition, hearing aids, and teleconferencing systems. However, speech enhancement still remains a challenging research topic. Many attempts to use noisy speech, training data, and physical properties of target sources have been undertaken for speech enhancement in the research, since we cannot directly know what the target speech source is.

Non-negative matrix factorization (NMF) [1] has attracted extensive attention due to its ease of feature detection in the acoustic frequency domain calculated with the short-time Fourier transform (STFT). Speech enhancement with the approach is conducted through the application of NMF to the amplitude spectrum in the acoustic frequency domain [2]. Here, a major shortcoming of NMF is that it cannot factorize any matrices other than non-negative matrices consisting of non-negative elements. This means that not the amplitude spectrum but the complex spectrum including the phase spectrum in the acoustic frequency domain cannot be applied to NMF. Therefore, NMF is not a suitable method to take phase information of speech into account although previous studies have suggested that not only amplitude but also phase information has played a significant role in speech enhancement [3–7]. A complex NMF framework [8] was presented by Kameoka *et al.* for the purpose of addressing the shortcomings in NMF and taking into account the phase spectrum in the acoustic frequency domain. King and Atlas simply called it "complex matrix factorization" (CMF) [9] and developed several additional algorithms in CMF [9–11].

Previous research, on the other hand, has demonstrated that spectral processing like spectral subtraction in the modulation frequency domain was a more effective scheme for speech enhancement than acoustic-domain schemes [12–15]. Non-negative tensor factorization (NTF) [16] was introduced for source separation [17] as a modulation domain approach that extended NMF. Moreover, Barker and Virtanen [18] proposed an algorithm for separating monaural audio signals by using the NTF of a modulation spectrogram [19], which meant the spectrogram of a modulation envelope. They suggested that the modulation spectrogram was able to represent redundant patterns across frequencies with similar features. The use of not only the amplitude but also the phase information is required in the modulation frequency domain to utilize more information on speech. However, existing methods based on matrix or tensor factorizations are not able to deal with any complex tensors such as complex spectrum including the phase spectrum.

This paper presents a novel algorithm of "complex tensor factorization (CTF)", which enables us to factorize an arbitrary complex tensor of rank 3. We propose a method of single-channel speech enhancement to factorize complex spectrum in the modulation frequency domain using CTF. The novelty of the proposed method is to estimate both separated amplitude and phase spectra in the acoustic frequency domain through factorizing complex spectra in the modulation frequency domain.

2. Background

2.1. Non-negative Matrix Factorization

NMF is an approximation algorithm for sparse representation to factorize an arbitrary non-negative matrix whose elements are non-negative into a product of two non-negative matrices. Its algorithm is based on solving an optimization problem typically expressed by [20]:

$$\text{Given} \quad : \quad \boldsymbol{X} \in \mathbb{R}^{\geq 0, L\times M},\ K \in \mathbb{N}^{+}, \tag{1}$$

$$\text{Factorize} \quad : \quad \boldsymbol{X} \simeq \boldsymbol{BW}, \tag{2}$$

$$\text{Minimize} \quad : \quad \sum_{l,m}\left| X_{l,m} - \sum_{k} B_{l,k}W_{k,m}\right|^{2}, \tag{3}$$

$$\text{Subject to} \quad : \quad \boldsymbol{B} \in \mathbb{R}^{\geq 0, L\times K},\ \boldsymbol{W} \in \mathbb{R}^{\geq 0, K\times M}, \tag{4}$$

where $\boldsymbol{X}$ is the non-negative matrix that corresponds to input data to be factorized by NMF. The K, $\boldsymbol{B}$, and $\boldsymbol{W}$ correspond to the number of bases, the base matrix, and the weight matrix. Here, the objective function in Eq. (3) is assumed to be the Euclidean distance, which is the simplest and most well known distance. The way the number of bases is determined remains a challenge in matrix factorizations.

Now, let us consider the application of NMF to single-channel speech enhancement. First, only noisy speech $y(t)$, where $y(t) = y_1(t) + y_2(t)$, is observed. Here, t is continuous time, $y_1(t)$ is clean speech, and $y_2(t)$ is noise or another signal. The complex spectrum, $\boldsymbol{Y}(\omega_I, \tau_I)$, of noisy speech in the acoustic frequency domain is given by STFT:

$$\boldsymbol{Y}(\omega_I, \tau_I) = \int_{-\infty}^{\infty} y(t) w_I(t - \tau_I) e^{-j\omega_I t} dt, \tag{5}$$

where ω_I, τ_I, and $w_I(t)$ indicate the frequency, time indices, and window function in STFT. The complex spectrum, $\boldsymbol{Y}(\omega_I, \tau_I) (= |\boldsymbol{Y}(\omega_I, \tau_I)| e^{j \arg \boldsymbol{Y}(\omega_I, \tau_I)})$, consists of its amplitude spectrum $|\boldsymbol{Y}(\omega_I, \tau_I)|$ and phase spectrum $\arg \boldsymbol{Y}(\omega_I, \tau_I)$.

A method of using the amplitude spectrum of noisy speech $|\boldsymbol{Y}(\omega_1, \tau_1)|$ as an input of non-negative matrix $\boldsymbol{X}$ for Eq. (1) in NMF has been known to be effective in speech enhancement. This method enables us to separate clean speech from noisy speech by using the base matrix calculated from the amplitude spectra of training data with NMF as the base matrix for noisy speech in NMF. We obtain the complex spectra of individual sources: $\hat{\boldsymbol{Y}}_i(\omega_I, \tau_I) = \hat{\boldsymbol{X}}_i(\omega_I, \tau_I) e^{j \arg \boldsymbol{Y}(\omega_I, \tau_I)}$ $(i = 1, 2)$, by using separated amplitude spectra $\hat{\boldsymbol{X}}_1(\omega_I, \tau_I)$ and $\hat{\boldsymbol{X}}_2(\omega_I, \tau_I)$. Here, we assume that the phase spectrum of each source is the same as that of noisy speech. Finally, conducting the inverse STFT (ISTFT) of the complex spectra of each source, i.e., $\hat{\boldsymbol{Y}}_1(\omega_I, \tau_I)$ and $\hat{\boldsymbol{Y}}_2(\omega_I, \tau_I)$, we estimate separated sources $\hat{y}_1(t)$ and $\hat{y}_2(t)$.

2.2. Complex non-negative Matrix Factorization

It has been suggested that CMF can overcome the non-negative constraint in NMF and factorize an arbitrary complex matrix into two non-negative matrices and a phase tensor [8]. The weight and base matrices are still non-negative. CMF can be defined as the following optimization problem:

$$\begin{aligned}
\text{Given} \quad &: \quad \boldsymbol{Y} \in \mathbb{C}^{L \times M},\ K \in \mathbb{N}^{+},\ \lambda \in \mathbb{R}, \\
&\quad\ g \in \mathbb{R} \mid 0 < g < 2, \qquad (6) \\
\text{Minimize} \quad &: \quad \frac{1}{2} \sum_{l,m} \left| Y_{l,m} - \sum_k B_{l,k} W_{k,m} e^{j\Phi_{l,k,m}} \right|^2 \\
&\quad\ + \lambda \sum_{k,m} |W_{k,m}|^g, \qquad (7) \\
\text{Subject to} \quad &: \quad \sum_l B_{l,k} = 1 \ \ (\forall k = 1, 2, \cdots, K), \\
&\quad\ \boldsymbol{B} \in \mathbb{R}^{\geq 0, L \times K}, \boldsymbol{W} \in \mathbb{R}^{\geq 0, K \times M}, \\
&\quad\ \boldsymbol{\Phi} \in \mathbb{R}^{L \times K \times M}, \qquad (8)
\end{aligned}$$

where $\boldsymbol{Y}$ and $\boldsymbol{\Phi}$ represent an input matrix for CMF and a phase tensor. The first and the second terms in the objective function of Eq. (7) correspond to the Frobenius norm and sparsity factor. λ and g are weighting and shape parameters to determine the sparsity factor. The $\sum_l B_{l,k} = 1$ in Eq. (8) was assumed to avoid indeterminacy in scaling. See Kameoka *et al.* [8] for the derivation of a solution to the optimization problem in Eqs. (6)-(8). An implementation of the solution can be found in King and Atlas [21].

CMF in speech enhancement is capable of directly factorizing the complex spectrum, $\boldsymbol{Y}(\omega_I, \tau_I)$, in Eq. (5) as input complex matrix $\boldsymbol{Y}$ in Eq. (6). Therefore, we can estimate separated sources, $\hat{y}_1(t)$ and $\hat{y}_2(t)$, by applying ISTFT to separated complex spectra, $\hat{\boldsymbol{Y}}_1(\omega_I, \tau_I)$ and $\hat{\boldsymbol{Y}}_2(\omega_I, \tau_I)$, that were obtained from CMF.

2.3. Modulation frequency domain

The acoustic frequency domain is used in speech enhancement based on matrix factorizations like NMF and CMF. However, it has been reported that modulation domain processing is a useful alternative to acoustic domain processing in the enhancement [12]. Moreover, Zhang and Zhao have proposed a method of subtraction for the amplitude spectrum in the modulation frequency domain [15]. In the modulation frequency domain, the complex spectra, $\boldsymbol{Z}_{\text{Re}}(\omega_I, \omega_{II}, \tau_{II})$ and $\boldsymbol{Z}_{\text{Im}}(\omega_I, \omega_{II}, \tau_{II})$, are calculated by applying the second STFT to real and imaginary parts of the complex spectrum in acoustic frequency domain, $\boldsymbol{Y}(\omega_I, \tau_I)$, of Eq. (5) as:

$$\boldsymbol{Z}_{\text{Re}} = \int_{-\infty}^{\infty} \text{Re}[\boldsymbol{Y}] w_{II}(\tau_I - \tau_{II}) e^{-j\omega_{II}\tau_I} d\tau_I, \tag{9}$$

$$\boldsymbol{Z}_{\text{Im}} = \int_{-\infty}^{\infty} \text{Im}[\boldsymbol{Y}] w_{II}(\tau_I - \tau_{II}) e^{-j\omega_{II}\tau_I} d\tau_I, \tag{10}$$

where ω_{II}, τ_{II}, and w_{II} indicate the modulation frequency, the time indices, and the window function in the second STFT. The $\boldsymbol{Z}_{\text{Re}}$ and $\boldsymbol{Z}_{\text{Im}}$ are complex tensors of rank 3. They showed that modulation domain processing played a larger role than the acoustic frequency phase under the conditions in which they were studied.

3. Proposed method

3.1. Complex spectra in modulation frequency domain

Our proposed method was aimed at applying novel tensor factorization to complex spectra, which are transformed by real, imaginary, and amplitude spectra in the acoustic frequency domain, in the modulation frequency domain for speech enhancement. It consists of three steps as we can see from the block diagram in Fig. 1:

1. Analysis stage, where complex spectra in the modulation frequency domain are extracted from noisy speech and training data by two STFTs.
2. Separation stage, where the mentioned spectra of noisy speech are separated by applying CTF, as explained in Subsection 3.2.
3. Synthesis stage by two ISTFTs for the separated spectra in the modulation frequency domain.

First, the complex spectrum, $\boldsymbol{Z}_{Am}(\omega_I, \omega_{II}, \tau_{II})$, transformed by amplitude spectrum in the acoustic frequency domain as well as Eqs. (9) and (10) is calculated in the analysis stage by:

$$\boldsymbol{Z}_{Am} = \int_{-\infty}^{\infty} |\boldsymbol{Y}| w_{II}(\tau_I - \tau_{II}) e^{-j\omega_{II}\tau_I} d\tau_I. \tag{11}$$

Six separated complex spectra in the modulation frequency domain, $\hat{\boldsymbol{Z}}_{\text{Re},1}$, $\hat{\boldsymbol{Z}}_{\text{Re},2}$, $\hat{\boldsymbol{Z}}_{\text{Im},1}$,, $\hat{\boldsymbol{Z}}_{\text{Im},2}$, $\boldsymbol{Z}_{Am,1}$, and $\boldsymbol{Z}_{Am,2}$, are

obtained in the separation stage. We calculated separated spectra, $\hat{\boldsymbol{Y}}_{\text{Re},1}$, $\hat{\boldsymbol{Y}}_{\text{Re},2}$, $\hat{\boldsymbol{Y}}_{\text{Im},1}$, $\hat{\boldsymbol{Y}}_{\text{Im},2}$, $\hat{\boldsymbol{Y}}_{Am,1}$, and $\hat{\boldsymbol{Y}}_{Am,2}$, in the acoustic frequency domain by using the ISTFT of the complex spectra in the modulation frequency domain. Combining the separated phase spectra with the separated amplitude spectra, we resynthesized the separated complex spectra, $\hat{\boldsymbol{Y}}_{Am,1}\exp[j\arg(\hat{\boldsymbol{Y}}_{\text{Re},1} + j\hat{\boldsymbol{Y}}_{\text{Im},1})] \equiv \hat{\boldsymbol{Y}}_1$ and $\hat{\boldsymbol{Y}}_{Am,2}\exp[j\arg(\hat{\boldsymbol{Y}}_{\text{Re},2} + j\hat{\boldsymbol{Y}}_{\text{Im},2})] \equiv \hat{\boldsymbol{Y}}_2$, in the acoustic frequency domain. Finally, the separated signals, $\hat{y}_1(t)$ and $\hat{y}_2(t)$, were computed with the second ISTFT of $\hat{\boldsymbol{Y}}_1$ and $\hat{\boldsymbol{Y}}_2$. The proposed method enabled us to factorize complex spectra in the modulation frequency domain through these steps.

3.2. Complex Tensor Factorization

Our proposed CTF to factorize an arbitrary complex tensor of rank 3 can now be written as:

$$\begin{aligned}
&\text{Given}: \boldsymbol{Z} \in \mathbb{C}^{L\times K'\times M},\ K \in \mathbb{N}^{+}, K' \in \mathbb{N}^{+},\ \lambda \in \mathbb{R}, \\
&\quad g \in \mathbb{R} \mid 0 < g < 2, \qquad (12) \\
&\text{Minimize}: \frac{1}{2}\sum_{l,m}\left|Z_{l,k',m} - \sum_{k} B_{l,k',k} W_{k,k',m} e^{j\Phi_{l,k,k',m}}\right|^2 \\
&\quad + \lambda \sum_{k,m} |W_{k,k',m}|^g \ \ (\forall k' = 1, 2, \cdots, K'), \qquad (13) \\
&\text{Subject to}: \sum_{l} B_{l,k',k} = 1 \ \ (\forall k = 1, 2, \cdots, K, \\
&\quad \forall k' = 1, 2, \cdots, K'),\ \boldsymbol{B} \in \mathbb{R}^{\geq 0, L\times K'\times K}, \\
&\quad \boldsymbol{W} \in \mathbb{R}^{\geq 0, K\times K'\times M}, \boldsymbol{\Phi} \in \mathbb{R}^{L\times K\times K'\times M}, \qquad (14)
\end{aligned}$$

where $\boldsymbol{Z}$ indicates an input tensor for CTF. The solution to the optimization problem of CTF is based on the solution to CMF by multiplicative update rule [8]; the update rules, $e^{j\tilde{\Phi}_{l,k,k',m}}$, $\tilde{B}_{l,k',k}$, and $W_{k,k',m}$ for the phase tensor, base tensor, and weight tensor are given by:

$$e^{j\tilde{\Phi}_{l,k,k',m}} = \frac{\bar{Z}_{l,k,k',m}}{|\bar{Z}_{l,k,k',m}|}, \qquad (15)$$

$$\tilde{B}_{l,k',k} = \frac{\sum_m W_{k,k',m}|\bar{Z}_{l,k,k',m}|/\beta_{l,k,k',m}}{\sum_m W^2_{k,k',m}/\beta_{l,k,k',m}}, \qquad (16)$$

$$\tilde{W}_{k,k',m} = \frac{\sum_l B_{l,k',k}|\bar{Z}_{l,k,k',m}|/\beta_{l,k,k',m}}{\sum_l \dfrac{B^2_{l,k',k}}{\beta_{l,k,k',m}} + \lambda g |W_{k,k',m}|^{g-2}}, \qquad (17)$$

where $F_{l,k',m} \equiv \sum_k B_{l,k',k} W_{k,k',m} e^{j\Phi_{l,k,k',m}}$, $\beta_{l,k,k',m} \equiv B_{l,k',k} W_{k,k',m} / \sum_n B_{l,k',n} W_{n,k',m}$, and $\bar{Z}_{l,k,k',m} \equiv B_{l,k',k} W_{k,k',m} e^{j\Phi_{l,k,k',m}} + \beta_{l,k,k',m}(Z_{l,k',m} - F_{l,k',m})$. In addition, we assumed an update rule for the phase tensor $P(\Phi)(\equiv e^{j\Phi_{l,k,k',m}})$ in the direction of k' to enhance speech more. Let us consider the input and output phase tensor in discrete number k' in its update, $P_{in}(\Phi)[k']$ and $P_{out}(\Phi)[k']$, and assume the following update rule for phase tensor:

$$P_{in}(\Phi)[k'] = \frac{\sum_u^{k'-1} P_{out}(\Phi)[u]}{k' - 1} \ \ (\forall k' = 2, 3, \cdots, K'). \qquad (18)$$

We assumed that initial base matrix has the values calculated by training data, initial weight matrix has random values, and initial phase tensor $P_{in}(\Phi)[k' = 1]$ is $Z_{l,1,m}/|Z_{l,1,m}|$.

CTF is applied to the complex spectra which are represented by Eqs. (9)-(11) as input complex tensors of rank 3. We assumed $L = N_{\omega_I}$, $M = N_{\tau_{II}}$, and

$$K = K' = N_{\omega_{II}}, \qquad (19)$$

in this algorithm for the number of elements and axis in tensors for Eqs. (12)-(14), where N_{ω_I}, $N_{\omega_{II}}$ and $N_{\tau_{II}}$ correspond to the number of frames for ω_I, ω_{II} and τ_{II} directions in each STFT. Note that Eq. (19) allows to automatically determine the number of bases, K and K'. The CTF algorithm is based on $N_{\omega_{II}}$ times CMFs from low to high modulation frequency, $k' = 1 \rightarrow 2 \rightarrow \cdots \rightarrow N_{\omega_{II}}$.

4. Evaluations

This section reports the results obtained from two kinds of experiments using speech sentences from the Texas Instruments Massachusetts Institute of Technology (TIMIT) database to evaluate the effectiveness of the proposed method in comparison with NMF and CMF. We investigated improvements to speech enhancement by measuring the signal to noise ratio loss (SNR loss) [22] and the signal to error ratio (SER) between clean $y_1(t)$ and restored speech $\hat{y}_1(t)$ from noisy speech $y(t)$. SER is defined as:

$$\text{SER}(y_1, \hat{y}_1) = 10\log_{10}\frac{\int_0^T (y_1(t))^2 dt}{\int_0^T (y_1(t) - \hat{y}_1(t))^2 dt}. \qquad (20)$$

The sampling frequency of all speech sentences used in the evaluation was 20 kHz.

The same parameters in all computations using CTF were applied, which did not depend on individual speech sentences. First, sparsity parameters, g and λ, in Eq. (13) were set to 1.0 and 10, which were based on parameter tests. The number of bases K and K' in Eq. (19) were determined by using $N_{\omega_{II}}$, which was calculated with the STFT window length and overlap ratio of CTF. STFT was computed using a Hanning window that was 64 ms long with 50% overlap, which corresponded to an $N_{\omega_{II}}$ of 33. The number of iterations in CTF was set to one. The initial phase tensors for the first CMF ($k' = 1$) in CTF were assumed to be phase tensors that were calculated by training data using CTF.

The sparsity parameters in Eq. (7) were set to $g = 1.0$ and $\lambda = 0.010$ for all computations with CMF, which were the same values as those in a previous study [10].

Five iterations in NMF and CMF were carried out for this evaluation in all computations because they seemed to be sufficient in terms of the convergence of objective functions in both NMF and CMF. We used 33 bases in NMF because the number of bases in CTF was also 33.

4.1. Experiment 1: Noisy speech

The first experiment evaluated the performance of the proposed approach for speech enhancement under a simple problem setting. We used noisy speech that consisted of clean speech sentences from the TIMIT database and white noise. The speech sentences were composed of two kinds of sentences uttered by 462 English female and male speakers. The first sentence was "She had your dark suit in greasy wash water all year" to estimate noisy speech. The second was "Don't ask me to carry an oily rag like that", which was used for training in NMF, CMF, and CTF. The SNRs between clean speech $y_1(t)$ and white noise data $y_2(t)$ were fixed from -10 to 20 dB at intervals of 10 dB.

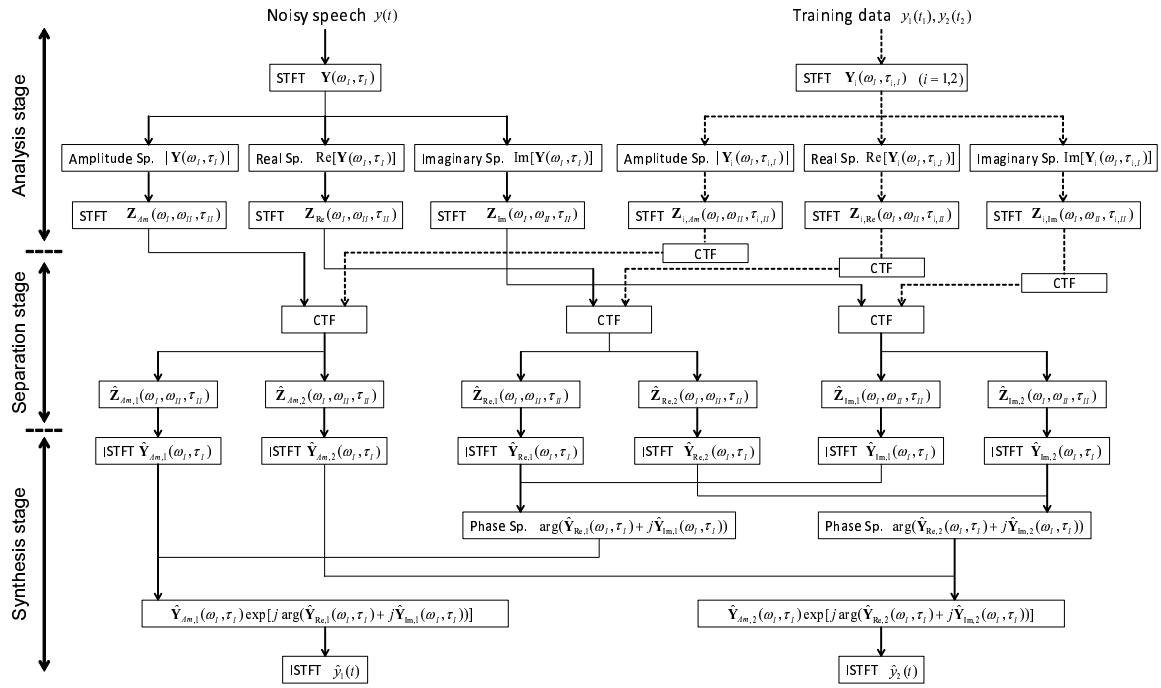

Figure 1: Block diagram of proposed method.

Table 1 summarized the average of improved SERs and SNR losses for the proposed method, NMF, and CMF with different white noise levels. We can see that CTF led to the best performance under all noise conditions. Thus, the proposed method enabled better speech enhancement during noisy speech than existing NMF and CMF.

Table 1: Comparison of averaged values of improved SER (ISER) and SNR loss in Experiment 1.

SNR	NMF		CMF		CTF	
	ISER	SNR loss	ISER	SNR loss	ISER	SNR loss
-10	10.8	0.913	11.6	0.914	14.0	0.896
0	7.98	0.834	8.39	0.844	9.12	0.794
10	2.75	0.726	4.25	0.734	4.51	0.660
20	-4.86	0.623	-1.75	0.634	0.626	0.507

4.2. Experiment 2: Mixed speech

A second experiment was conducted to apply the proposed method to mixed speech uttered by females and males. There were a total of 50 mixed speeches uttered by 50 females and 50 males from the TIMIT database. One mixed speech consisted of two different speech sentences uttered by a female and a male. Different speech sentences for training and estimation were used in this experiment. The SNRs between one clean speech $y_1(t)$ and another clean speech $y_2(t)$, which were regarded as noise in this experiment, were fixed from -10 to 20 dB at intervals of 10 dB.

Table 2 lists the average of improved SERs for the proposed method, NMF, and CMF with different sound levels for other speech. We can see that CTF led to the best performance for all sound levels. Thus, the proposed method enabled better speech enhancement than existing NMF and CMF. In addition, the results suggested that there is a fair chance of achieving improvements in speech enhancement by using factorizations in the modulation frequency domain rather than in the acoustic frequency domain.

Table 2: Comparison of averaged values of ISER and SNR loss in Experiment 2.

SNR	NMF		CMF		CTF	
	ISER	SNR loss	ISER	SNR loss	ISER	SNR loss
-10	6.47	0.910	11.0	0.922	13.9	0.902
0	3.39	0.897	8.22	0.854	9.30	0.801
10	-3.93	0.898	4.53	0.746	4.62	0.678
20	-13.4	0.906	-1.12	0.657	0.907	0.539

5. Conclusions

This paper proposed a novel method of speech enhancement based on tensor factorization in the modulation frequency domain by expanding CMF. The results from our experiments revealed that the proposed method outperformed the existing acoustic domain methods based on NMF and CMF in terms of improvements to the restoration of waveforms under the conditions in which we did the studies. This performance can be regarded as the effect of factorization in the modulation frequency domain by not only taking into account amplitude but also phase information. We intend to further investigate expanding the algorithm by taking into account speech dereverberation in future work and the selection of optimum transforms other than STFT to express the characteristics of signals with tensors.

6. Acknowledgements

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and by Secom Science and Technology Foundation.

7. References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects with non-negative matrix factorization," *Nature*, 401, pp. 788–791, 1999.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for music transcription," *Proc. WASPAA*, pp. 177–180, 2003.

[3] B. J. Shannon and K. K. Paliwal, "Role of phase estimation in speech enhancement," *Proc. Interspeech2006-ICSLP*, Pittsburgh, Pennsylvania, pp. 1427–1430, 2006.

[4] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening test," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[5] K. K. Paliwal, K. Wojcicki, and B. J. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[6] M. Unoki and M. Akagi, "A Method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, pp. 261–279, 1999.

[7] N. Nower, Y. Liu, and M. Unoki, "Restoration of instantaneous amplitude and phase using Kalman filter for speech enhancement," *Proc. ICASSP2014*, pp. 4633–4637, 2014.

[8] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," *Proc. ICASSP2009*, pp. 3437–3440, 2009.

[9] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," *Proc. ICASSP2010*, pp. 4206–4209, 2010.

[10] B. King and L. Atlas, "Single-channel source separation using complex matrix factorization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2591–2597, 2011.

[11] B. King, "New methods of complex matrix factorization for single-channel source separation and analysis," *Ph.D. thesis, University of Washington*, 2012.

[12] K. K. Paliwal, K. Wojcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, pp. 450–475, 2010.

[13] K. K. Paliwal, B. Schwerin, and K. Wojcicki, "Role of modulation magnitude and phase spectrum towards speech intelligibility," *Speech Communication*, vol. 53, pp. 327–339, 2011.

[14] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Communication*, vol. 53, pp. 818–829, 2011.

[15] Y. Zhang and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Communication*, vol. 55, pp. 509–522, 2013.

[16] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," *Proc. International Conference of Machine Learning*, 2005.

[17] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," *Proc. Irish Signals and Systems Conference*, 2005.

[18] T. Barker and T. Virtanen, "Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation," *Proc. Interspeech2013*, pp. 827–831, 2013.

[19] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," *Proc. ICASSP1997*, pp. 1647–1650, 1997.

[20] J. Bronson and P. Depalle, "Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources," *Proc. ICASSP2014*, pp. 7525–7529, 2014.

[21] B. King and L. Atlas, "Complex matrix factorization toolbox version 1.0 for MATLAB," `https://sites.google.com/a/uw.edu/isdl/projects/cmf-toolbox`, *University of Washington*, September 2012.

[22] J. Ma and P. C. Loizou, "SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Communication*, vol. 53, pp. 304–359, 2011.