MISSING COMPONENT RESTORATION FOR MASKED SPEECH SIGNALS BASED ON TIME-DOMAIN SPECTROGRAM FACTORIZATION

Shogo Seki$^{1)}$, *Hirokazu Kameoka*$^{2)}$, *Tomoki Toda*$^{3)}$, *Kazuya Takeda*$^{4)}$

$^{1)}$Graduate School of Informatics, Nagoya University
$^{2)}$NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
$^{3)}$Information Technology Center, Nagoya University
$^{4)}$Institutes of Innovation for Future Society, Nagoya University

ABSTRACT

While time-frequency masking is a powerful approach for speech enhancement in terms of signal recovery accuracy (e.g., signal-to-noise ratio), it can over-suppress and damage speech components, leading to limited performance of succeeding speech processing systems. To overcome this shortcoming, this paper proposes a method to restore missing components of time-frequency masked speech spectrograms based on direct estimation of a time domain signal. The proposed method allows us to take account of the local interdependencies of the elements of the complex spectrogram derived from the redundancy of a time-frequency representation as well as the global structure of the magnitude spectrogram. The effectiveness of the proposed method is demonstrated through experimental evaluation, using spectrograms filtered with masks to enhance of noisy speech. Experimental results show that the proposed method significantly outperformed conventional methods, and has the potential to estimate both phase and magnitude spectra simultaneously and precisely.

***Index Terms*—** Missing component restoration, Time-domain spectrogram factorization

1. INTRODUCTION

The presence of background noise can significantly degrade the quality of speech traveling through transmission systems and negatively affect the performance of speech recognition and speech conversion systems. The performance of these systems can be improved by suppressing the noise in observed audio signals and enhancing the target speech.

One effective approach for speech enhancement involves time-frequency masking, which extracts only the components in the time-frequency slots that are expected to be dominated by the target speech [1]. There are several ways to perform time-frequency masking. For example, by using microphone inputs we can cluster time-frequency slots according to the direction of arrival of each source [2]. For monaural recording, we can use deep neural networks to assign a source label to every time-frequency slot, or partition the spectrogram into different source regions according to the local "texture" of the spectrogram [3, 4]. While these methods allow aggressive suppression of noise components, they can also over-suppress the speech component and damage its acoustic features. As a result, the performance of speech processing systems can be limited, even if a high signal-to-noise ratio (SNR) is obtained. To overcome this limitation, this paper deals with the problem of restoring the missing components of over-masked spectrograms.

One conventional missing component restoration approach for masked spectrograms is based on Non-negative Matrix Factorization (NMF) [5, 6]. NMF-based methods attempt to restore missing components by assuming that the entire spectrogram can be approximated as a low-rank matrix, namely, as the product of two non-negative matrices [7]. Modeling the entire spectrogram in this way amounts to assuming that the magnitude spectrum observed at each time frame can be approximated as the sum of a limited number of spectral templates. The signal can then be reconstructed using a phase reconstruction algorithm [8]. Since spectrograms are generally redundant representations of time-domain signals, the magnitude and phase of each time-frequency slot are in fact interdependent on each other. In other words, spectrograms must satisfy a certain constraint in order to be associated with time-domain signals. [8] uses this fact as the basis for devising a phase reconstruction algorithm. This implies that we can also use this relationship as a clue to help restore the missing components of masked spectrograms. However, the performance of common phase reconstruction methods is still insufficient, and NMF-based methods also require some prior information about the target speech to be effective.

Recently, a time-domain extension of NMF called Time-domain Spectrogram Factorization (TSF) has been proposed [9]. As the name implies, TSF performs NMF-like signal decomposition in the time domain by taking account of the intrinsically redundant structure of spectrograms. While regular NMF approximates an observed magnitude spectrogram into the sum of rank-1 spectrograms, TSF decomposes an observed time-domain signal into the sum of L signal components, such that the magnitude spectrogram of each component is as close to a rank-1 structure as possible. In this work, we propose using TSF to directly estimate the waveform signal such that its magnitude spectrogram can be approximated as a low-rank matrix so that missing component restoration and phase reconstruction can be performed jointly in a principled manner.

Cepstral Distance Regularization (CDR) is a recently proposed technique used in semi-supervised NMF (SSNMF), which aims to enhance target speech in both the spectral and cepstral domains [10]. CDR does this by optimizing a combined objective function composed of an NMF-based model fitting criterion defined in the spectral domain and a Gaussian mixture model-based probability distribution defined in the Mel-Frequency Cepstral Coefficient (MFCC) domain.

This paper proposes a TSF-based missing component restoration method which combines the conventional methods discussed above in a novel manner. The proposed method considers; 1) cues

978-1-5090-6341-3/17/$31.00 ©2017 IEEE

of local dependencies of each component, which are detected using redundancy in time-frequency domain expression, and/or 2) prior information about the target speech in a feature space, in addition to cues considered by conventional NMF-based methods. The effectiveness of the proposed method is demonstrated through the experimental restoration of masked speech spectrograms which are obtained by applying Ideal Binary Mask (IBM) filters to noisy speech. The restoration performance of the proposed TSF-based method is then compared with that of conventional NMF-based methods.

2. PRELIMINARIES

2.1. Non-negative Matrix Factorization (NMF)

NMF can be used to approximate an observed magnitude spectrogram, interpreted as a non-negative matrix $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{K \times M}$, as a low-rank matrix by factorizing X into the product of two non-negative matrices $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{K \times L}$ and $\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{L \times M}$:

$$\boldsymbol{X} \approx \boldsymbol{H}\boldsymbol{U}. \tag{1}$$

This amounts to assuming that the magnitude spectrum observed at each time frame can be approximated as the sum of L basis spectra:

$$X_{k,m} \simeq \hat{X}_{k,m} = \sum_l H_{k,l} U_{l,m}. \tag{2}$$

If the magnitude spectrogram of an audio signal of interest can be assumed to have a low-rank structure, missing components in the magnitude spectrogram can be restored by fitting the NMF model (1) over the observable regions [7]. The time-domain signal can then be synthesized for example by using a phase reconstruction technique [8].

2.2. Time-domain Spectrogram Factorization (TSF)

TSF is a novel signal decomposition technique that aims to directly decompose an observed time-domain signal $\boldsymbol{s} \in \mathbb{R}^N$ (where N denotes the number of the samples of the entire signal) into the sum of L signal components:

$$\boldsymbol{s} = \sum_l \boldsymbol{s}_l, \tag{3}$$

such that the magnitude spectrogram of $\boldsymbol{s}_l$ becomes as close to a rank-1 (or low-rank) structure as possible. This idea can be formulated as an optimization problem of minimizing:

$$\mathcal{I}(\boldsymbol{\theta}) = \sum_l \sum_{k,m} (|\boldsymbol{\psi}_{k,m}^{\mathsf{H}} \boldsymbol{s}_l| - H_{k,l} U_{l,m})^2 + \mathcal{R}(\boldsymbol{U}), \tag{4}$$

$$\text{subject to } \sum_l \boldsymbol{s}_l = \boldsymbol{s}, \tag{5}$$

where $\mathcal{R}(\boldsymbol{U})$ is a sparse regularization term. $\boldsymbol{\psi}_{k,m}^{\mathsf{H}} \boldsymbol{s}_l$ represents the time-frequency element of $\boldsymbol{s}_l$ (i.e., a Short-Time Fourier Transform (STFT), or Wavelet Transformation), and $\boldsymbol{\psi}_{k,m} \in \mathbb{C}^N$ is a complex sinusoid windowed at time frame t_m with center frequency ω_k.

Since this method allows to directly estimate the signal components $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L$ in the time domain, the phase reconstruction procedure is implicitly involved in the spectrogram factorization process. This gives TSF its name.

2.3. Cepstral Distance Regularization

Although speech enhancement methods based on semi-supervised NMF are powerful in terms of signal recovery accuracy (e.g., signal-to-noise ratio), they do not necessarily lead to an improvement in the quality of the enhanced speech in the feature domain [11]. To overcome this limitation, CDR has recently proposed to jointly enhance speech both in the spectral and cepstral domains [10]. CDR forces the estimated spectrogram to follow the same statistical distribution as the training data in the feature space by introducing a regularization term given as the negative logarithm of the Gaussian mixture density:

$$-\mathcal{K}\left(\hat{\boldsymbol{\mathcal{X}}}\right) = -\log \prod_m \sum_p w_p \prod_q \mathcal{N}\left(\hat{\mathcal{X}}_{q,m}; \mu_{p,q}, \sigma_{p,q}^2\right), \tag{6}$$

$$\hat{\mathcal{X}}_{q,m} = \sum_r c_{q,r} \log \sum_k f_{r,k} \hat{X}_{k,m}, \tag{7}$$

where $\hat{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{Q \times M}$ is a sequence of MFCCs of $\hat{\boldsymbol{X}}$. $\boldsymbol{f} = \{f_{r,k}\}_{r,k} \in \mathbb{R}^{R \times K}$ and $\boldsymbol{c} = \{c_{q,r}\}_{q,r} \in \mathbb{R}^{Q \times R}$ represent a mel-filterbank coefficient matrix and an inverse discrete cosine transform matrix, respectively. Eq. (6) represents the negative log-likelihood of a Gaussian Mixture Model (GMM) with parameters $\{w_p, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p\}_p$. Note that the GMM parameters are pretrained using training examples of clean speech.

3. PROPOSED METHOD

3.1. Problem Setting

Let $\boldsymbol{Y} \in \mathbb{C}^{K \times M}$ be an observed complex spectrogram with missing components whose components are represented as $Y_{k,m}$ and where $k \in \{1, \ldots, K\}$ and $m \in \{1, \ldots, M\}$ are indices of frequency bins, and time frames, respectively. By using Γ to denote the set of the observable time-frequency slots of $\boldsymbol{Y}$, here we assume that the STFT coefficients in the missing regions are zero:

$$Y_{k,m} = 0, \quad ((k,m) \notin \Gamma). \tag{8}$$

We would like to estimate these components so that the time-domain signal can be reconstructed.

3.2. Objective Function Design

We can use the TSF framework to impute missing components by considering local interdependencies of the elements of a complex spectrogram. We can also borrow the idea from the conventional NMF-based approach to estimate the magnitude part of the missing components by assuming the magnitude spectrogram to have a low-rank structure. Additionally, we use CDR to ensure that the restored spectrogram follows a pretrained distribution in the cepstral domain. Hence, we propose introducing the following objective function to be minimized:

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) = &\sum_{(k,m) \in \Gamma} |\boldsymbol{\psi}_{k,m}^{\mathsf{H}} \boldsymbol{s} - Y_{k,m}|^2 \\ &+ \lambda_1 \sum_{k,m} \mathcal{D}.\left(|\boldsymbol{\psi}_{k,m}^{\mathsf{H}} \boldsymbol{s}| \mid \hat{X}_{k,m}\right) \\ &+ \lambda_2 \sum_{(k,m) \in \Gamma} \mathcal{D}.\left(|Y_{k,m}| \mid \hat{X}_{k,m}\right) \\ &- \lambda_3 \mathcal{K}\left(\hat{\boldsymbol{\mathcal{X}}}\right), \end{aligned} \tag{9}$$

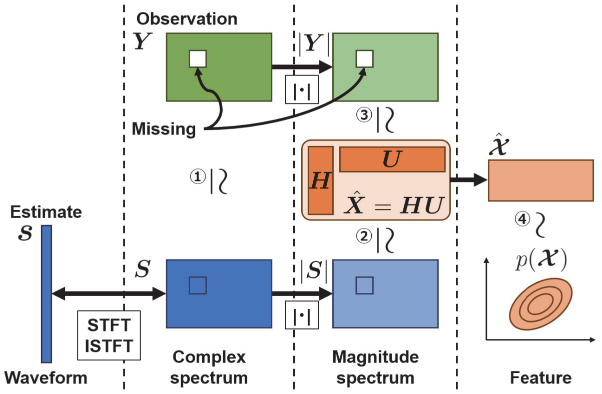


Fig. 1. Illustration of the designed objective function

where λ_1, λ_2 and λ_3 are hyperparameters that weigh the importance of the second, third and fourth terms, respectively, $\boldsymbol{\theta} = \{\boldsymbol{s}, \boldsymbol{H}, \boldsymbol{U}\}$ is the set of parameters to be optimized, and $\mathcal{D}.$ is a divergence measure between non-negative arguments. In this paper, either squared Euclidean distance or Kullback-Leibler (KL) divergence are used. Fig. 1 shows an overview of the proposed method, where the circled numbers correspond the individual terms of the objective function. In Eq. (9), the first term represents the squared error between an observed complex spectrogram and that of the estimated signal $\boldsymbol{s}$ over the observable regions Γ. It is important to note that $\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}$ always satisfies the condition that all complex spectrograms must satisfy and thus the redundancy of the time-frequency representation is implicitly considered. The second term represents the error between Eq. (2) and the magnitude spectrogram of $\boldsymbol{s}$, which connects the first and the third term effects. The third term represents the error between Eq. (2) and the observed magnitude spectrogram. This term corresponds to the objective function of the conventional NMF-based approach. The fourth term is cepstral distance regularization term given in Eq. (6). This forces $\hat{\boldsymbol{X}}$ to follow the statistical distribution of target speech in the feature space domain. By optimizing the objective function, the missing components of observed spectrogram $\boldsymbol{Y}$ can be restored by satisfying the provided constraints.

4. PARAMETER ESTIMATION ALGORITHM

Here we describe a novel convergence-guaranteed algorithm for minimizing Eq. (9) based on a majorization-minimization (MM) principle.

4.1. Majorization-minimization principle

We use $F(\theta)$ to denote an objective function that we want to minimize with respect to θ. $F^+(\theta, \alpha)$ is defined as a "majorizer" for $F(\theta)$ if it satisfies:

$$F(\theta) = \min_{\alpha} F^+(\theta, \alpha). \tag{10}$$

We call α an auxiliary variable. By using $F^+(\theta)$, $F(\theta)$ can be iteratively decreased according to the following theorem:

Lemma 1
$F(\theta)$ is non-increasing under the updates, $\theta \leftarrow \mathrm{argmin}_\theta F^+(\theta, \alpha)$ and $\alpha \leftarrow \mathrm{argmin}_\alpha F^+(\theta, \alpha)$.

4.2. Update rules for $\boldsymbol{s}$

In Eq. (9), the first and second terms are related to parameter $\boldsymbol{s}$. When $\mathcal{D}.$ is defined as the squared Euclidean distance, as in [9], we can show:

$$\begin{aligned}
&\sum_{k,m} \mathcal{D}_{\mathrm{EU}} \left(|\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}| \mid \hat{X}_{k,m} \right) \\
&= \sum_{k,m} \left[|\boldsymbol{\psi}^{H}_{k,m}\boldsymbol{s}|^2 - 2|\boldsymbol{\psi}^{H}_{k,m}\boldsymbol{s}|\hat{X}_{k,m} + \hat{X}^2_{k,m} \right] \\
&\leq \sum_{k,m} |\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s} - \hat{X}_{k,m} a_{k,m}|^2.
\end{aligned} \tag{11}$$

Here, we can use the right-hand side of this inequality as a majorizer for the second term of Eq. (9), where $\boldsymbol{a} = \{a_{k,m}\}_{k,m}$ is an auxiliary parameter. The equality holds when:

$$a_{k,m} = \frac{\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}}{|\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}|}. \tag{12}$$

When $\mathcal{D}.$ is defined as the KL-divergence, we can show:

$$\begin{aligned}
&\sum_{k,m} \mathcal{D}_{\mathrm{KL}} \left(|\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}| \mid \hat{X}_{k,m} \right) \\
&= \sum_{k,m} \left[|\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}| \log \frac{|\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}|}{\hat{X}_{k,m}} - |\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}| + \hat{X}_{k,m} \right] \\
&\leq \sum_{k,m} \left[F_{k,m}|\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}|^2 - 2\mathrm{Re} \left[G^*_{k,m}\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s} \right] + \hat{X}_{k,m} \right] + \mathrm{const.},
\end{aligned} \tag{13}$$

where $D_{k,m}$, $F_{k,m}$, and $G_{k,m}$ are given by:

$$D_{k,m} = \log \frac{\xi_{k,m}}{\hat{X}_{k,m}} - 2, \tag{14}$$

$$F_{k,m} = \begin{cases} \frac{D_{k,m}}{2b_{k,m}} + \frac{1}{\xi_{k,m}}, & (D_{k,m} \geq 0) \\ \frac{1}{\xi_{k,m}}, & (D_{k,m} < 0) \end{cases}, \tag{15}$$

$$G_{k,m} = \begin{cases} 0, & (D_{k,m} \geq 0) \\ -D_{k,m}a^*_{k,m}/2, & (D_{k,m} < 0) \end{cases}. \tag{16}$$

Similarly, we can use the right-hand side of this inequality as a majorizer for the case of the KL-divegence where $\boldsymbol{b} = \{b_{k,m}\}_{k,m}$, and $\boldsymbol{\xi} = \{\xi_{k,m}\}_{k,m}$ are auxiliary parameters. The equality of Eq. (13) satisfies when:

$$b_{k,m} = \xi_{k,m} = |\boldsymbol{\psi}^{\mathsf{H}}_{k,m}\boldsymbol{s}|. \tag{17}$$

Since both Eqs. (11) and (13) are differentiable and convex, an optimal update for s minimizing Eq. (11) or Eq. (13) can be found using gradient methods. In the case of the squared Euclidean distance, parameter $\boldsymbol{s}$ can be efficiently updated in the following way. For the first term of Eq. (9), let $\tilde{Y}_{k,m}$ defined as follows:

$$\tilde{Y}_{k,m} = \begin{cases} Y_{k,m}, & ((k,m) \in \Gamma) \\ S_{k,m}, & ((k,m) \notin \Gamma) \end{cases}, \tag{18}$$

Since $\sum_{(k,m)\notin\Gamma} |\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s} - S_{k,m}|^2 \geq 0$, we obtain:

$$\begin{aligned}&\sum_{(k,m)\in\Gamma} |\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s} - Y_{k,m}|^2 \\ \leq &\sum_{(k,m)\in\Gamma} |\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s} - Y_{k,m}|^2 + \sum_{(k,m)\notin\Gamma} |\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s} - S_{k,m}|^2 \\ = &\sum_{k,m} |\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s} - \tilde{Y}_{k,m}|^2. \end{aligned} \tag{19}$$

Thus, we can also use the right-hand side of Eq. (19) as a majorizer where $\boldsymbol{S} = \{S_{k,m}\}_{k,m}$ is an additional set of auxiliary parameters. The equality of Eq. (18) holds when:

$$S_{k,m} = \boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s}. \tag{20}$$

Since this majorizer is given as a quadratic function of $\boldsymbol{s}$, obtain an update rule for s analytically as follows:

$$\begin{aligned}\boldsymbol{s} =&\frac{1}{1+\lambda_1}\left(\sum_{k,m}\mathrm{Re}[\boldsymbol{\psi}_{k,m}\boldsymbol{\psi}_{k,m}^{\mathsf{H}}]\right)^{-1} \\ &\times\left(\sum_{k,m}\mathrm{Re}[\boldsymbol{\psi}_{k,m}(\tilde{Y}_{k,m} + \lambda_1\hat{X}_{k,m}a_{k,m})]\right). \end{aligned} \tag{21}$$

Although Eq. (21) contains inverse matrix computation, this can be avoided by selecting $\boldsymbol{\psi}_{k,m}$, so that $\sum_{k,m}\boldsymbol{\psi}_{k,m}\boldsymbol{\psi}_{k,m}^{\mathsf{H}}$ becomes a circulant matrix. It can be diagonalized using discrete Fourier transform matrix $\boldsymbol{F}$ as follows: $\sum_{k,m}\mathrm{Re}[\boldsymbol{\psi}_{k,m}\boldsymbol{\psi}_{k,m}^{\mathsf{H}}] = \boldsymbol{F}\boldsymbol{V}\boldsymbol{F}^{\mathsf{H}}$. The inverse matrix can now be calculated efficiently. For example, when $\boldsymbol{\psi}_{k,m}$ represents an STFT with a square-root Hanning window, diagonal matrix $\boldsymbol{V}$ becomes an identity matrix.

When $\mathcal{D}.$ is defined as the a KL-divergence, the inverse matrix computation is unavoidable because the matrix to be inverted does not become a circulant matrix. Instead of trying to obtain an update rule with an analytical form, here we choose to update $\boldsymbol{s}$ be obtained using a gradient method where the gradient is given in the form:

$$\begin{aligned}\nabla_{\boldsymbol{s}}\mathcal{I}(\boldsymbol{\theta}) =&2\sum_{k,m}\mathrm{Re}[\boldsymbol{\psi}_{k,m}\{(r_{k,m}+\lambda_1F_{k,m})\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s} \\ &-(r_{k,m}Y_{k,m}+\lambda_1G_{k,m})\}]. \end{aligned} \tag{22}$$

Here, $r_{k,m}$ is binary variable defined as:

$$r_{k,m} = \begin{cases}1, & ((k,m)\in\Gamma) \\ 0, & ((k,m)\notin\Gamma)\end{cases}. \tag{23}$$

Note that terms with the form $\sum_{k,m}\mathrm{Re}[\boldsymbol{\psi}_{k,m}\cdot]$ can be computed efficiently using the Fast Fourier Transform (FFT).

4.3. Update rules for $\boldsymbol{H}$ and $\boldsymbol{U}$

In Eq. (9), the second, third and fourth terms are related to parameters $\boldsymbol{H}$, and $\boldsymbol{U}$. When $\mathcal{D}.$ is defined as the squared Euclidean distance, the update rules for $\boldsymbol{H}$, and $\boldsymbol{U}$ can be obtained in the same manner as the regular NMF, as follows:

$$H_{k,l} = \frac{\sum_m(\lambda_1|\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s}| + \lambda_2 r_{k,m}|Y_{k,m}|)U_{l,m}}{\sum_m(\lambda_1+\lambda_2 r_{k,m})\frac{U_{l,m}^2}{\beta_{k,l,m}}}, \tag{24}$$

$$U_{l,m} = \frac{\sum_k(\lambda_1|\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s}| + \lambda_2 r_{k,m}|Y_{k,m}|)H_{k,l}}{\sum_k(\lambda_1+\lambda_2 r_{k,m})\frac{H_{k,l}^2}{\beta_{k,l,m}}}, \tag{25}$$

where $\boldsymbol{\beta} = \{\beta_{k,l,m}\}_{k,l,m}$ satisfies:

$$\beta_{k,l,m} = \frac{H_{k,l}U_{l,m}}{\hat{X}_{k,m}}. \tag{26}$$

When $\mathcal{D}.$ is defined as the KL-divergence, the update rules for $\boldsymbol{H}$ and $\boldsymbol{U}$ can be derived as in [10], as follows:

$$H_{k,l} = \frac{-\mathsf{b}_{k,l} + \sqrt{\mathsf{b}_{k,l}^2 - 4\mathsf{a}_{k,l}\mathsf{c}_{k,l}}}{2\mathsf{a}_{k,l}}, \tag{27}$$

$$U_{l,m} = \frac{-\mathsf{e}_{l,m} + \sqrt{\mathsf{e}_{l,m}^2 - 4\mathsf{d}_{l,m}\mathsf{f}_{l,m}}}{2\mathsf{d}_{l,m}}, \tag{28}$$

where $\mathsf{a}_{k,l}$, $\mathsf{b}_{k,l}$, $\mathsf{c}_{k,l}$, $\mathsf{d}_{l,m}$, $\mathsf{e}_{l,m}$, and $\mathsf{f}_{l,m}$ are defined as follows:

$$\begin{aligned}\mathsf{a}_{k,l} =&\sum_m(\lambda_1+\lambda_2 r_{k,m})U_{l,m} \\ &+\lambda_3\sum_{r,m}\left(A_{r,m}p(\zeta_{r,m}) + \frac{\delta_{B_{r,m}\geq 0}|B_{r,m}|}{\phi_{r,m}}\right)f_{r,k}U_{l,m}, \end{aligned} \tag{29}$$

$$\begin{aligned}\mathsf{b}_{k,l} =&-\sum_m(\lambda_1|\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s}| + \lambda_2 r_{k,m}|Y_{k,m}|)\beta_{k,l,m} \\ &-\lambda_3\sum_{r,m}\delta_{B_{r,m}<0}|B_{r,m}|v_{r,k,l,m}, \end{aligned} \tag{30}$$

$$\mathsf{c}_{k,l} = -\lambda_3\sum_{r,m}A_{r,m}\frac{\rho_{r,k,l,m}^2}{f_{r,k}U_{l,m}},$$

$$\begin{aligned}\mathsf{d}_{l,m} =&\sum_k(\lambda_1+\lambda_2 r_{k,m})H_{k,l} \\ &+\lambda_3\sum_{r,k}\left(A_{r,m}p(\zeta_{r,m}) + \frac{\delta_{B_{r,m}\geq 0}|B_{r,m}|}{\phi_{r,m}}\right)f_{r,k}H_{k,l}, \end{aligned} \tag{31}$$

$$\begin{aligned}\mathsf{e}_{l,m} =&-\sum_k(\lambda_1|\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s}| + \lambda_2 r_{k,m}|Y_{k,m}|)\beta_{k,l,m} \\ &-\lambda_3\sum_{r,k}\delta_{B_{r,m}<0}|B_{r,m}|v_{r,k,l,m}, \end{aligned} \tag{32}$$

$$\mathsf{f}_{l,m} = -\lambda_3\sum_{r,k}A_{r,m}\frac{\rho_{r,k,l,m}^2}{f_{r,k}H_{k,l}}. \tag{33}$$

δ_x is an indicator function which takes the value of one when condition x is satisfied, otherwise its value is zero. Note that $L_{r,m}$, $A_{r,m}$, $B_{r,m}$, and $p(\cdot)$ are defined as follows:

$$L_{r,m} = \sum_k f_{r,k}\hat{X}_{k,m}, \tag{34}$$

$$A_{r,m} = \sum_{p,q}\frac{\eta_{p,m}c_{q,r}^2}{2\sigma_{p,q}^2\omega_{p,q,r,m}}, \tag{35}$$

$$B_{r,m} = -\sum_{p,q}\frac{\eta_{p,m}c_{q,r}\varphi_{p,q,r,m}}{\sigma_{p,q}^2\omega_{p,q,r,m}}, \tag{36}$$

$$p(\zeta_{r,m}) = \frac{2\log\zeta_{r,m}}{\zeta_{r,m}} + \frac{1}{\zeta_{r,m}^2}. \tag{37}$$

$\boldsymbol{\eta} = \{\eta_{p,m}\}_{p,m}$, $\boldsymbol{\varphi} = \{\varphi_{p,q,r,m}\}_{p,q,r,m}$, $\boldsymbol{\rho} = \{\rho_{r,k,l,m}\}_{r,k,l,m}$, $\boldsymbol{v} = \{v_{r,k,l,m}\}_{r,k,l,m}$, $\boldsymbol{\zeta} = \{\zeta_{r,m}\}_{r,m}$, and $\boldsymbol{\phi} = \{\phi_{r,m}\}_{r,m}$ are all auxiliary parameters satisfying following relations:

$$\eta_{p,m} = \frac{w_p \prod_q \mathcal{N}(\mathcal{X}_{q,m}; \mu_{p,q}, \sigma^2_{p,q})}{\sum_{p'} w_{p'} \prod_{q'} \mathcal{N}(\mathcal{X}_{q',m}; \mu_{p',q'}, \sigma^2_{p',q'})}, \tag{38}$$

$$\varphi_{p,q,r,m} = c_{q,r} \log L_{r,m} + \omega_{p,q,r,m}(\mu_{p,q} - \mathcal{X}_{q,m}), \tag{39}$$

$$\rho_{r,k,l,m} = v_{r,k,l,m} = \frac{f_{r,k} H_{k,l} U_{l,m}}{\sum_{k',l'} f_{r,k'} H_{k',l'} U_{l',m}}, \tag{40}$$

$$\zeta_{r,m} = \phi_{r,m} = L_{r,m}, \tag{41}$$

and $\boldsymbol{\omega} = \{\omega_{p,q,r,m}\}_{p,q,r,m}$ is an arbitrary positive constant parameter satisfying $\sum_r \omega_{p,q,r,m} = 1$.

5. EXPERIMENTAL EVALUATIONS

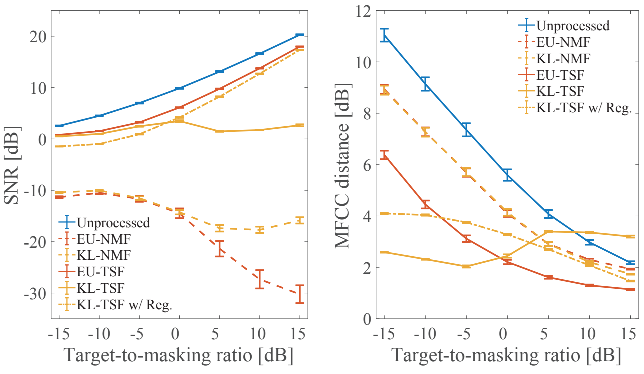
5.1. Settings

The performance of the proposed methods was evaluated through experiments with speech spectrograms which had been masked using IBMs for noise elimination. The masked spectrograms were prepared using IBMs constructed of clean speech and noise data, which can be represented as:

$$M_{\mathrm{IBM}} = \begin{cases} 1, & \left(10 \log_{10} \frac{|S^{(\mathrm{C})}_{k,m}|^2}{|S^{(\mathrm{N})}_{k,m}|^2} > \epsilon\right) \\ 0, & (\text{otherwise}) \end{cases}, \tag{42}$$

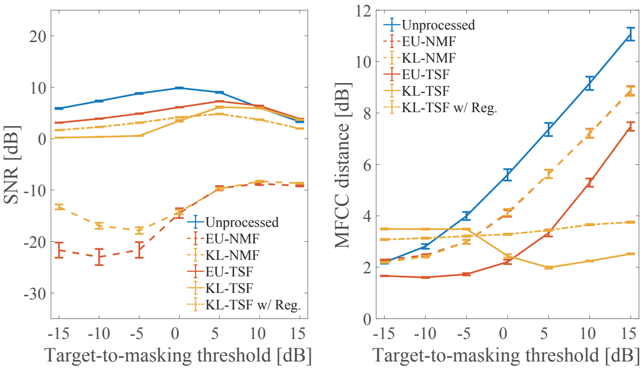
where $S^{(\mathrm{C})}_{k,m}$ and $S^{(\mathrm{N})}_{k,m}$ are complex spectrograms of clean speech and noise, respectively, and ϵ is the threshold determining whether each component activates or not. As clean data, 200 utterances of 20 speakers from ATR 503 database, including males and females were used [12]. Babble noise was added to each clean speech sample at various SNR or threshold settings in Eq. (42) while building the IBMs for noisy speech, and then the spectrograms of the noisy speech were masked. Three datasets were built for the experiments; *Target-to-masking ratio dataset (TMR dataset)* in which noisy speech were made under varying SNR conditions, while the corresponding IBMs were constructed at a fixed threshold parameter (0 dB), *Target-to-masking threshold dataset (TMT dataset)* in which noisy speech were made at a fixed SNR setting (0 dB), but the corresponding IBMs were constructed at varying thresholds. and a *Over-masking dataset* in which noisy speech were made at a fixed SNR setting (0 dB), and the corresponding IBMs were constructed at a fixed threshold parameter (0 dB), however, the existing components through IBM filtering were moreover erased by making them zeros randomly and compulsorily, and it was considered in more practical conditions.

Three TSF-based methods were investigated as proposed methods: a TSF using the squared Euclidean distance (EU-TSF), and the TSFs using KL-divergence with or without cepstral distance regularization (KL-TSF w/ Reg., KL-TSF w/o Reg). Two NMF-based methods using the squared Euclidean distance (EU-NMF) and the KL-divergence (KL-NMF) were used to represent conventional methods. Each speech signal was sampled at 16 kHz, and the spectrograms were obtained through frame analysis using 32 ms and 16 ms shifts with square-root Hanning windows. The total number of basis spectra was set to 30, and the total number of iterations for parameter updating was 200. Weight parameters λ_1, λ_2, λ_3 were adjusted during the first half of the iterations so that each term of the objective function in Eq. (9) had the same magnitude, and these weight parameters were then fixed during the second half of the iterations. For cepstral distance regularization, 0-to-13th MFCCs with 20-dimensional mel-filterbanks were extracted from 100 other utterances of the individual speakers and were for GMM training. The mixture component of the GMM was set to 30. For KL-TSF, the Adadelta technique was used for gradient descent for $\boldsymbol{s}$ [13]. A phase spectrogram for the NMF-based methods was reconstructed using the Griffin-Lim algorithm with 100 iterations [8]. As measurements of performance, SNRs and the MFCC distances between the restored speech and the corresponding clean speech were used. In addition, restoration of each masked spectrogram was repeated 3 times owing to the weakening effect of the initial values, and measurements were averaged over all of the iterations and speech.



(a) *SNR* (b) *MFCC distance*

Fig. 2. Results for *TMR dataset*



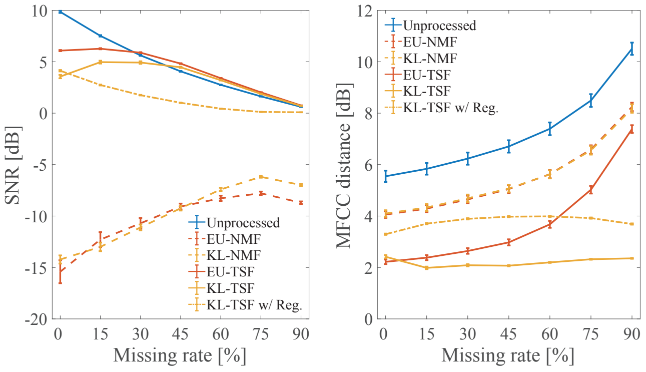
(a) *SNR* (b) *MFCC distance*

Fig. 3. Results for *TMT dataset*

5.2. Results

Figs. 2–4 show the SNR and the MFCC distance results for *TMR dataset*, *TMT dataset*, and *Over-masking dataset*. In Fig. 2, the horizontal axis shows SNR settings, while in Fig. 3 it represents the threshold settings of the IBMs. In Fig. 4, the horizontal axis shows missing rate for existing components of IBMs. The vertical axes in these figures represent performance. Error bars in the figures represent 95 % confidence intervals. Unprocessed results, whose waveform signals were obtained by reproducing the masked spectrograms straightforwardly without any reconstruction methods, are also shown in each figure.

These results show that the proposed TSF-based methods outperformed conventional NMF-based methods. Especially, we can see that the SNRs of TSF-based methods follow similar tendencies



(a) *SNR*

(b) *MFCC distance*

Fig. 4. Results for *Over-masking dataset*

to that of the unprocessed result unlike NMF-based methods. This is because that the NMF-based methods estimate not missing correct phase information, but estimate consistent phase information for the reconstructed spectrogram, which leads insufficient performance.

Moreover, the TSF-based methods maintained high SNRs similar to Unprocessed while greatly improving the MFCC distances. The proposed method using the squared Euclidean distance (EU-TSF) delivered especially stable results. In the over-suppress conditions (Fig. 4), we can see that the proposed methods using either the squared Euclidean distance (EU-TSF) or the KL-divergence (KL-TSF) exceed unprocessed results both in the SNRs and the MFCC distances. This suggests that the proposed methods are potential to restore the missing components well by applying over-masked spectrograms, and especially KL-TSF could achieve quite robust performance for the spectrograms.

These experiments also show that cepstral distance regularization does not consistently improve restoration performance. This is because the intensity of regularization can be unsuitable and can actually degrade performance. The balancing of error functions and regularization terms has not been sufficiently researched and remains a challenging problem.

6. CONCLUSION

This paper proposed a novel missing component restoration method for masked speech spectrograms based on a TSF signal decomposition model. The proposed method attempts to utilize as many acoustical cues as possible, e.g., cues observed in spectrograms as well as cues from spectrograms of target speech in a feature space, and, if possible, to directly estimate waveform signals. The experimental results showed that the proposed TSF-based restoration significantly outperform conventional NMF-based methods, and has potential to estimate both magnitude and phase spectra simultaneously and precisely. These results also demonstrated that a part of the TSF-based restoration methods has quite robust performance for over-suppressing occurring in time-frequency masking.

7. ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Number 17H01763.

8. REFERENCES

[1] Nathalie Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, pp. 126–137, 1999.

[2] Ozgur Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Pprocessing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[3] Yuxuan Wang and DeLiang Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[4] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[5] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[6] Paris Smaragdis and Judith C Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* IEEE, 2003, pp. 177–180.

[7] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Missing data imputation for time-frequency representations of audio signals," *Journal of Signal Processing Systems*, vol. 65, no. 3, pp. 361–370, 2011.

[8] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[9] Hirokazu Kameoka, "Multi-resolution signal decomposition with time-domain spectrogram factorization," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 86–90.

[10] Li Li, Hirokazu Kameoka, Takuya Higuchi, and Hiroshi Saruwatari, "Semi-supervised joint enhancement of spectral and cepstral sequences of noisy speech," *Interspeech 2016*, pp. 3753–3757, 2016.

[11] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.

[12] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[13] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.