MEL-GENERALIZED CEPSTRAL REGULARIZATION FOR DISCRIMINATIVE NON-NEGATIVE MATRIX FACTORIZATION

Li Li[1], *Hirokazu Kameoka*[2] *and Shoji Makino*[1]

[1]University of Tsukuba, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan

ABSTRACT

The non-negative matrix factorization (NMF) approach has shown to work reasonably well for monaural speech enhancement tasks. This paper proposes addressing two shortcomings of the original NMF approach: (1) the objective functions for the basis training and separation (Wiener filtering) are inconsistent (the basis spectra are not trained so that the separated signal becomes optimal); (2) minimizing spectral divergence measures does not necessarily lead to an enhancement in the feature domain (e.g., cepstral domain) or in terms of perceived quality. To address the first shortcoming, we have previously proposed an algorithm for Discriminative NMF (DNMF), which optimizes the same objective for basis training and separation. To address the second shortcoming, we have previously introduced novel frameworks called the cepstral distance regularized NMF (CDRNMF) and mel-generalized cepstral distance regularized NMF (MGCRNMF), which aim to enhance speech both in the spectral domain and feature domain. This paper proposes combining the goals of DNMF and MGCRNMF by incorporating the MGC regularizer into the DNMF objective function and proposes an algorithm for parameter estimation. The experimental results revealed that the proposed method outperformed the baseline approaches.

Index Terms— Discriminative non-negative matrix factorization, mel-generalized cepstral representation, speech enhancement, single-channel

1. INTRODUCTION

Speech enhancement is a technique for recovering a speech signal underlying an observed noisy speech signal. Since the presence of noise can significantly degrade the quality of speech transmission systems and the performance of such applications as speech recognition and speech conversion, many methods have been proposed over the decades.

For monaural speech enhancement tasks, the non-negative matrix factorization (NMF) approach [1, 2] has attracted a lot of attention after being proposed as a powerful approach. In recent years, deep neural network-based approaches have shown to be significantly effective particularly for supervised speech enhancement tasks [3, 4]. However, the NMF approach still remains attractive under unsupervised or semi-supervised settings or when only a limited amount of training data is available. The basic idea behind the NMF approach is that approximating the magnitude or power spectrum observed at each time frame as a linear sum of basis spectra scaled by time-varying amplitudes amounts to factorizing the spectrogram of an observed signal, interpreted as a non-negative matrix, into the product of two non-negative matrices. In a supervised/semi-supervised setting, NMF is first applied to the spectrograms of training instances to train the basis spectra of speech (and noise). At test time, NMF is applied to the spectrogram of a test mixture signal, where each subset of the basis spectra is fixed at the pretrained spectra. The underlying speech components can then be separated out using a Wiener filter constructed by the estimated power spectrograms of speech and noise. Although this approach has shown to work reasonably well, there are two shortcomings to be addressed: (1) the basis spectra obtained in the conventional way do not ensure that the separated signal will be optimal at test time due to the inconsistency between the objective functions for training and separation (Wiener filtering); (2) NMF does not necessarily lead to an enhancement in the feature domain (e.g., cepstral domain) or in terms of perceived quality, which implies naively using NMF as a front-end processing for e.g., speech recognition and speech transmission systems does not always lead to satisfactory results.

To address the first shortcoming, we have previously focused on a framework called the Discriminative NMF (DNMF) [5], which provides a way to train the basis spectra so that the output of the Wiener filter becomes as close to the spectrogram of each training example as possible. While the convergence of the basis training algorithm proposed in the original work of DNMF [5] is not guaranteed, we have proposed a convergence-guaranteed algorithm based on the majorization-minimization principle [6]. To overcome the second shortcoming, we have previously proposed two extensions of NMF, namely, cepstral distance regularized NMF (CDRNMF) [7] and NMF with mel-generalized cepstral regularization (MGCRNMF) [8]. These methods have

This work was supported by JSPS KAKENHI Grant Number 17H01763 and SECOM Science and Technology Foundation.

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

allowed us to jointly enhance speech in both the spectral and feature domains by optimizing a combined objective function of an NMF-based model-fitting criterion defined in the spectral domain and a GMM-based probability distribution defined in the feature domain. The MGCRNMF framework uses the mel-generalized cepstral (MGC) representation [9] of speech as the feature, which is widely used in parametric speech synthesis. Since the MGC representation is an auditory-motivated representation of speech spectra, using this representation as the feature has shown to contribute to enhancing the perceived quality of enhanced speech.

This paper proposes a novel approach combining the goals of DNMF and MGCRNMF by incorporating the MGC regularizer into the DNMF objective function and proposes an algorithm for the parameter estimation.

The remaining part of the paper proceeds as follows: we briefly review the conventional NMF approach for speech enhancement in subsec. 2.1 and the established methods DNMF and MGCRNMF in the rest of sec. 2. In sec. 3, we introduce the proposed method and derive the update rules for parameter estimation algorithm based on majorization-minimization principle. In the experimental section (sec. 4), we define the data set, investigate the hyperparameters of the model and compare the proposed method with the established methods. Conclusions are given in sec. 5.

2. CONVENTIONAL METHODS

2.1. NMF for speech enhancement

Given an observed power spectrogram of a noisy speech signal $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$, where ω and t are frequency and time indices, NMF factorizes it into the product of a non-negative basis matrix $\boldsymbol{W} = [\boldsymbol{W}^s, \boldsymbol{W}^n]$ and a non-negative coefficient (activition) matrix $\boldsymbol{H} = [\boldsymbol{H}^s; \boldsymbol{H}^n]$, where $\boldsymbol{W}^l = (W^l_{\omega,k})_{\Omega\times K^l} \in \mathbb{R}^{\geq 0,\Omega\times K^l}$ with $l = \{s, n\}$ is pretrained by applying NMF to the spectrograms of training samples $\boldsymbol{R}^l = (R^l_{\omega,t})_{\Omega\times T}$. A typical criterion for this is

$$\boldsymbol{W}^l = \underset{\boldsymbol{W}^l}{\operatorname{argmin}}\, \mathcal{D}(\boldsymbol{R}^l|\boldsymbol{W}^l\boldsymbol{H}^l) \tag{1}$$

with a column-wise normalization of $\boldsymbol{W}^l$, where $\mathcal{D}$ is a cost function that measures the difference between $\boldsymbol{R}^l$ and $\boldsymbol{W}^l\boldsymbol{H}^l$. At test time, $\boldsymbol{W}$ is fixed at the trained basis spectra and $\boldsymbol{H}$ is the variable to be estimated so that the objective function

$$\boldsymbol{H} = \underset{\boldsymbol{H}}{\operatorname{argmin}}\, \mathcal{D}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H}), \tag{2}$$

is minimized subject to non-negativity. NMF leads to different optimization problems according to the definition of the cost function $\mathcal{D}$. Here we consider the generalized Kullback Leibler (KL) divergence as a goodness-of-fit criterion.

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H})$$
$$= \sum_{\omega,t}\left(Y_{\omega,t}\log\frac{Y_{\omega,t}}{[\boldsymbol{W}\boldsymbol{H}]_{\omega,t}} - Y_{\omega,t} + [\boldsymbol{W}\boldsymbol{H}]_{\omega,t}\right), \tag{3}$$

where $[\cdot]_{i,j}$ denotes the $\{i,j\}$-th element of a matrix. Once the underlying speech components $X^s_{\omega,t} = [\boldsymbol{W}^s\boldsymbol{H}^s]_{\omega,t}$ and noise components $X^n_{\omega,t} = [\boldsymbol{W}^n\boldsymbol{H}^n]_{\omega,t}$ are estimated, the enhanced speech can be separated out using the Wiener filter constructed with the estimated power spectrogram of speech and noise

$$\hat{\boldsymbol{X}}^s = \frac{\boldsymbol{W}^s\boldsymbol{H}^s}{\boldsymbol{W}\boldsymbol{H}} \otimes \boldsymbol{Y}, \tag{4}$$

where $\frac{\cdot}{\cdot}$ and $\otimes$ here are element-wise operations.

2.2. Discriminative NMF

To eliminate the inconsistency between the objective functions for training (1) and separation (4), instead of applying NMF to speech and noise training samples individually to train the basis spectra, Weninger [5] proposed directly using the reconstruction error of the separated signals as an objective function for the basis training

$$\boldsymbol{W} = \underset{\boldsymbol{W}}{\operatorname{argmin}}\, \mathcal{D}_{KL}\left(\boldsymbol{R}^s \middle| \frac{\boldsymbol{W}^s\boldsymbol{H}^s}{\boldsymbol{W}\boldsymbol{H}} \otimes \boldsymbol{M}\right), \tag{5}$$
$$\text{subject to} \quad \forall k,\ \sum_{\omega} W_{\omega,k} = 1,$$

where $\boldsymbol{M}$ denotes the spectrograms of the created mixture signals by adding training speech data to noise data. This framework is called discriminative NMF (DNMF) by analogy with the discriminative models for classification or regression.

2.3. Mel-generalized cepstral regularization

We first introduce the mel-generalized cepstral (MGC) representation [9] which plays a key role in the method proposed in [8]. MGC representaion is a parametric model for spectral envelope of speech with frequency resolution similar to the human auditory systems, which is described by $M + 1$ coefficients and two hyperparameters γ and α,

$$\mu_\omega = l_\gamma^{-1}\left(\sum_{m=0}^{M} c(m)\Psi_\alpha^m(e^{j\omega})\right) \tag{6}$$
$$= \begin{cases} \left(1 + \gamma \sum\limits_{m=0}^{M} c(m)\Psi_\alpha^m(e^{j\omega})\right)^{1/\gamma} & (0 < |\gamma| \leq 1) \\ \exp \sum\limits_{m=0}^{M} c(m)\Psi_\alpha^m(e^{j\omega}) & (\gamma = 0) \end{cases}.$$

The function $l_\gamma^{-1}(\cdot)$ is the inverse of the generalized logarithmic function

$$l_\gamma(\omega) = \begin{cases} (\omega^\gamma - 1)/\gamma & (0 < |\gamma| \leq 1) \\ \log\omega & (\gamma = 0) \end{cases}, \tag{7}$$

parameterized by γ. $\Psi_\alpha(z)$ is an all-pass function given by

$$\Psi_\alpha(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \tag{8}$$

which can be seen as a frequency warping function parameterized by $|\alpha| < 1$. The coefficients $\boldsymbol{c} = [c(0), \ldots, c(M)]^\mathsf{T}$ are called the MGC coefficients (MGCCs). Given a spectral sequence, the MGCCs are defined as the inverse Fourier transform of the generalized logarithmic power spectrum calculated on a warped frequency scale. Note that MGC representation takes the all-pole spectral model and the cepstral representation as special cases when $(\gamma, \alpha) = (-1, 0)$ and $(\gamma, \alpha) = (0, 0)$ respectively. When the sampling frequency is 16 kHz, the phase characteristic of the all-pass function becomes a good approximation to the mel scale with $\alpha = 0.42$ and to the bark scale with $\alpha = 0.55$ [11].

With the motivation to ensure that the features of the underlying speech components $X^s_{\omega,t}$ in cepstral domain are also enhanced, by utilizing a codebook consisting of I prototype spectra μ_i represented using MGC representation which is pretrained by K-means algorithm using clean speech samples, we have previously proposed mel-generalized cepstral distance regularized NMF (MGCRNMF) combining a NMF-based model-fitting criterion (2) and a penalty term defined as the Itakura-Saito (IS) divergence [10] between $X^s_{\omega,t} = [\boldsymbol{W}^s \boldsymbol{H}^s]_{\omega,t}$ and a prototype spectrum $S_{\omega,t}(\theta)$ selected from the pretrained codebook

$$\mathcal{J}(\boldsymbol{W}^s, \boldsymbol{H}^s, \boldsymbol{\theta}) = \sum_{\omega,t} \Big(\frac{X^s_{\omega,t}}{S_{\omega,t}(\theta)} - \log \frac{X^s_{\omega,t}}{S_{\omega,t}(\theta)} - 1 \Big), \tag{9}$$

where $S_{\omega,t}(\theta) = \beta_{t,r_t} \mu_{\omega,r_t}$. Here, $\theta = \{r_t, \beta_{t,r_t}\}$ consists of a set of cluster indicator variables $r_t \in \{1, \ldots, I\}$ describing to which of the I clusters the t-th speech spectrum is assigned and the correspongding scaling parameters β_{t,r_t} introduced to eliminate the scaling indeterminacy. At test time, we find the prototype spectrum μ_i closest to $\boldsymbol{X}^s_t = [X^s_{1,t}, \ldots, X^s_{\Omega,t}]^T$ in terms of the IS divergence and the $S_{\omega,t}(\theta)$ can be easily obtained by multiplying the selected μ_i to the optimal scaling

$$\hat{\beta}_{t,i} = \frac{1}{\Omega} \sum_\omega \frac{X^s_{\omega,t}}{\mu_{\omega,i}}. \tag{10}$$

Note that the less (9) becomes, the more simliar speech envelope $\boldsymbol{X}^s$ has to the clean speech one, which means the features in cepstral domain are enhanced.

3. PROPOSED METHOD

In this section, we introduce a novel approach which combines the goals of DNMF and MGCRNMF by incorporating the MGC regularizer into the DNMF objective function. Then we derive an computationally efficient algorithm for parameter estimation.

Specifically, the proposed method uses DNMF for basis training and considers an optimization problem combining a NMF-based model-fitting criterion (2) and a novel MGC regularizer obtained by replacing $X^s_{\omega,t}$ by $\hat{X}^s_{\omega,t}$,

$$\tilde{\mathcal{J}}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta}) = \sum_{\omega,t} \Big(\frac{\hat{X}^s_{\omega,t}}{S_{\omega,t}(\theta)} - \log \frac{\hat{X}^s_{\omega,t}}{S_{\omega,t}(\theta)} - 1 \Big), \tag{11}$$

where $\hat{X}^s_{\omega,t}$ are the enhanced speech spectra obtained by Wiener filtering (4). With the new regularization term, the problem we are interested in can be cast as

$$\begin{aligned} \boldsymbol{H} &= \operatorname*{argmin}_{\boldsymbol{H}} \mathcal{F}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta}), \\ \mathcal{F}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta}) &= \mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H}) + \lambda \tilde{\mathcal{J}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\theta}) \end{aligned} \tag{12}$$

with $\boldsymbol{W}$ is fixed at the pretrained spectral basis using (5). $\lambda \geq 0$ here is a weight parameter to measure the importance of the regularization term relative to the NMF cost.

3.1. Majorization-minimization principle

Although minimizing the objective function including the regularization term (11) directly is analytically difficult, we can derive a computationally efficient algorithm to find a locally optimal solution based on majorization-minimization (MM) principle [12, 13].

Suppose $F(\Theta)$ is an objective function that we wish to minimize with respect to Θ. Majorization-minimization principle considers to construct a “majorizer” $F^+(\Theta, \alpha)$ defined as a function satisfying $F(\Theta) = \min_\alpha F^+(\Theta, \alpha)$, where α is called an auxiliary parameter. An algorithm that consists of iteratively minimizing $F^+(\Theta, \alpha)$ with respect to Θ and α is guaranteed to converge to a stationary point of the objective function. It should be noted that this concept is adopted in many existing algorithms [1, 14].

3.2. Update rules and algorithm

Here, we derive a majorizer for the objective function (12) with respect to $\boldsymbol{H}^s$ and $\boldsymbol{H}^n$ when target MGC representation $S_{\omega,t}(\hat{\theta})$ with θ fixed to $\hat{\theta}$. First, $\mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H})$ involves a “log-of-sum” form of $W_{k,\omega} H_{k,t}$. Since the negative logarithm function is a convex function, we can invoke Jensen's inequality to construct an upper bound of $\mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H})$ having a “sum-of-logs” form in the same way as [1]

$$\begin{aligned} &\mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H}) \leq \mathcal{D}^+_{KL}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H}) \\ &\mathcal{D}^+_{KL}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H}) \\ &\qquad \overset{c}{=} \sum_{\omega,t} \Bigg(- Y_{\omega,t} \sum_k \zeta_{k,\omega,t} \log \frac{W_{k,\omega} H_{k,t}}{\zeta_{k,\omega,t}} + X_{\omega,t} \Bigg), \end{aligned} \tag{13}$$

where $=^c$ denotes equality up to a constant term and $\zeta_{k,\omega,t}$ is a positive weight that sums to unity, $\sum_k \zeta_{k,\omega,t} = 1$. It can be shown that equality of (13) holds if and only if

$$\zeta_{k,\omega,t} = \frac{W_{k,\omega}H_{k,t}}{\sum_{k'=1}^{K} W_{k',\omega}H_{k',t}}. \tag{14}$$

Then, we focus on the regularization term

$$\tilde{\mathcal{J}}(\boldsymbol{W},\boldsymbol{H};\hat{\boldsymbol{\theta}}) \overset{c}{=} \sum_{\omega,t} \Big(\frac{Y_{\omega,t}G^s_{\omega,t}}{S_{\omega,t}(\hat{\theta})G_{\omega,t}} - \log G^s_{\omega,t} + \log G_{\omega,t} \Big), \tag{15}$$

where $G^s_{\omega,t} = \sum_{k=1}^{K_s} W^s_{k,\omega}H^s_{k,t}$ and $G_{\omega,t} = \sum_{k=1}^{K} W_{k,\omega}H_{k,t}$. To construct an upper bound for the first term of (15), we can invoke the Lemma 1 introduced in [6]

$$\frac{G^s_{\omega,t}}{G_{\omega,t}} \le \frac{\tau_{\omega,t}{G^s_{\omega,t}}^2}{2} + \frac{1}{2\tau_{\omega,t}G^2_{\omega,t}}. \tag{16}$$

The equality of (16) holds if and only if

$$\tau_{\omega,t} = \frac{1}{G^s_{\omega,t}G_{\omega,t}}. \tag{17}$$

Since a quadratic function is convex, we can apply Jensen's inequality to ${G^s_{\omega,t}}^2$, which yields

$${G^s_{\omega,t}}^2 \le \sum_{k=1}^{K_s} \frac{{W^s_{\omega,k}}^2{H^s_{k,t}}^2}{\alpha_{k,\omega,t}}, \tag{18}$$

where $\alpha_{k,\omega,t} > 0$ is also a positive number that sums to unity, i.e., $\sum_k \alpha_{k,\omega,t} = 1$. The equality of (18) holds if and only if

$$\alpha_{k,\omega,t} = \frac{W^s_{k,\omega}H^s_{k,t}}{\sum_{k'=1}^{K_s} W^s_{k',\omega}H^s_{k',t}}. \tag{19}$$

We can use the fact that $1/x^2$ is convex in the first quadrant and use Jensen's inequality to obtain a majorizer:

$$\frac{1}{G^2_{\omega,t}} \le \sum_{k=1}^{K} \frac{\xi^3_{k,\omega,t}}{W^2_{k,\omega}H^2_{k,t}}, \tag{20}$$

where $\xi_{k,\omega,t} > 0$ and $\sum_k \xi_{k,\omega,t} = 1$. It can be proved that the equality of this inequality holds if and only if

$$\xi_{k,\omega,t} = \frac{W_{k,\omega}H_{k,t}}{\sum_{k'=1}^{K} W_{k',\omega}H_{k,t'}}. \tag{21}$$

By substituting (18) and (20) into (16), the majorizer for the first term can be written as

$$\frac{G^s_{\omega,t}}{G_{\omega,t}} \le \sum_{k=1}^{K_s} \frac{\tau_{\omega,t}{W^s_{k,\omega}}^2{H^s_{k,t}}^2}{2\alpha_{k,\omega,t}} + \sum_{k=1}^{K} \frac{\xi^3_{k,\omega,t}}{2\tau_{\omega,t}W^2_{k,\omega}H^2_{k,t}}. \tag{22}$$

As regards the second term, Jensen's inequality can be invoked again since $-\log G^s_{\omega,t}$ is convex in $G^s_{\omega,t}$,

$$-\log G^s_{\omega,t} \le -\sum_{k=1}^{K_s} \gamma_{k,\omega,t} \log \frac{W^s_{k,\omega}H^s_{k,t}}{\gamma_{k,\omega,t}}, \tag{23}$$

where $\gamma_{k,\omega,t}$ is a positive weight that sums to unity. The equality of (23) holds if and only if

$$\gamma_{k,\omega,t} = \frac{W^s_{k,\omega}H^s_{k,t}}{\sum_{k'=1}^{K_s} W^s_{k',\omega}H^s_{k',t}}. \tag{24}$$

The third term $\log G_{\omega,t}$ is concave in $G_{\omega,t}$. Hence, we can use the fact that a tangent line to the graph of a differentiable concave function lies entirely above the graph:

$$\log G_{\omega,t} \le \sum_{k=1}^{K} \frac{W_{k,\omega}H_{k,t}}{\eta_{\omega,t}} + \log \eta_{\omega,t} - 1, \tag{25}$$

where $\eta_{\omega,t}$ is an arbitrary positive number. The equality of this inequality holds if and only if

$$\eta_{\omega,t} = G_{\omega,t}. \tag{26}$$

From (22), (23) and (25), we can construct a majorizer for the regularization term as

$$\begin{aligned}
&\tilde{\mathcal{J}}(\boldsymbol{W},\boldsymbol{H};\hat{\boldsymbol{\theta}}) \le \tilde{\mathcal{J}}^+(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\Gamma};\hat{\boldsymbol{\theta}})\\
&= \sum_{k,\omega,t} \frac{\tau_{\omega,t}Y_{\omega,t}{W^s_{k,\omega}}^2{H^s_{k,t}}^2}{2\alpha_{k,\omega,t}S_{\omega,t}(\hat{\theta})} + \sum_{k,\omega,t} \frac{\xi^3_{k,\omega,t}Y_{\omega,t}}{2\tau_{\omega,t}S_{\omega,t}(\hat{\theta})W^2_{k,\omega}H^2_{k,t}}\\
&- \sum_{k,\omega,t} \gamma_{k,\omega,t} \log \frac{W^s_{k,\omega}H^s_{k,t}}{\gamma_{k,\omega,t}} + \sum_{k,\omega,t} \frac{W_{k,\omega}H_{k,t}}{\eta_{\omega,t}} + d,
\end{aligned}$$

where $\boldsymbol{\Gamma} = \{\zeta_{k,\omega,t}, \tau_{\omega,t}, \gamma_{k,\omega,t}, \eta_{\omega,t}, \alpha_{k,\omega,t}, \xi_{k,\omega,t}\}$ denotes a set of all the auxiliary variables and d denotes a constant term. The upper bound for the objective function can be easily obtained by combining the majorizers for each term as

$$\mathcal{F}^+(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\Gamma};\hat{\boldsymbol{\theta}}) = \mathcal{D}^+_{KL}(\boldsymbol{Y}|\boldsymbol{W}\boldsymbol{H}) + \lambda\tilde{\mathcal{J}}^+(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\Gamma};\hat{\boldsymbol{\theta}}).$$

The update rules for $H_{k,t}$ can be obtained by setting at zeros the partial derivatives of the derived majorizer with respect to $H^s_{k,t}$ and $H^n_{k,t}$. Thus, the update rules can be obtained as the positive solution of the following quartic and cubic equations:

$$\begin{aligned}
&\sum_{\omega} \frac{\tau_{\omega,t}Y_{\omega,t}}{2\alpha_{k,\omega,t}S_{\omega,t}(\hat{\theta})}{W^s_{k,\omega}}^2{H^s_{k,t}}^4 + \sum_{\omega} \frac{W^s_{k,\omega}}{\eta_{\omega,t}}{H^s_{k,t}}^3\\
&\quad - \sum_{\omega} \gamma_{k,\omega,t}{H^s_{k,t}}^2 - \sum_{\omega} \frac{Y_{\omega,t}\xi^3_{k,\omega,t}}{2\tau_{\omega,t}S_{\omega,t}(\hat{\theta}){W^s_{k,\omega}}^2} = 0,
\end{aligned} \tag{27}$$

$$\sum_{\omega} \frac{W^n_{k,\omega}}{\eta_{\omega,t}}{H^n_{k,t}}^3 - \sum_{\omega} \frac{Y_{\omega,t}\xi^3_{k,\omega,t}}{2\tau_{\omega,t}S_{\omega,t}(\hat{\theta}){W^n_{k,\omega}}^2} = 0. \tag{28}$$

It is noteworthy that all the parameters can be updated in parallel using these update rules, which means this algorithm is well suited to parallel implementations. Furthermore, since each of the update rules consists of a negative 0th-order term and a negative 2nd-order term, it turns out that there is only one positive solution, implying that there is no need to solve a solution selection problem.

Algorithm. 1 shows the whole procedure.

Algorithm 1 Algorithm presented in subsec. 3.2

```
Require: pretrained speech basis $\boldsymbol{W}$ and $I$ MGC prototypes
    $\boldsymbol{\mu}$, parameters $\lambda$ and $MaxIter$
random initialize $\boldsymbol{H}^s$ and $\boldsymbol{H}^n$
for $iter = 1$ to $MaxIter$ do
    if $iter \leq 50$ then
        update $\boldsymbol{H}^s$ and $\boldsymbol{H}^n$ using SNMF
    else
        calculate the enhanced speech $\hat{\boldsymbol{X}}^s$ using (4)
        for Frame $t = 1$ to $T$ do
            $r_t = \arg\min_i \mathcal{D}_{IS}(\boldsymbol{X}^s_t, \boldsymbol{\mu}_i)$
            compute $\beta_{t,i}$ using (10)
            $S_{\omega,t}(\hat{\theta}) = \beta_{t,r_t}\mu_{\omega,r_t}$
        end for
        update auxiliary variables $\boldsymbol{\Gamma}$ using (14), (17), (19),
        (21) , (24) and (26)
        update $\boldsymbol{H}^s$, $\boldsymbol{H}^n$ by solving the equations (27) and (28)
    end if
end for
```

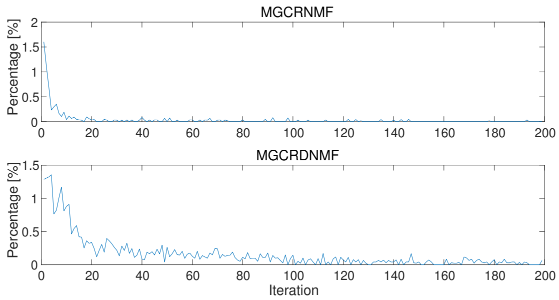

Fig. 1. Percentage of the number of frames which change the cluster to be assigned during the updating of MGCRNMF (upper) and MGCRDNMF (bottom). The average number of 10 samples randomly selected under 5 noise types are shown in the figure.

Table 1. A comparison of runtime [sec] between updating indicator variables at each iteration and updating them only at the first iteration using MGCRNMF and the proposed method. The length of the test data was 5 seconds.

	w/ update	w/o update
MGCRNMF	135.5534	1.9471
MGCRDNMF	259.9060	126.9996

4. EXPERIMENTS

To evaluate the effect of the proposed method for speech enhancement task, we tested supervised NMF (SNMF) [2], Discriminative NMF (DNMF) [6], NMF with mel-generalized cepstral regularization (MGCRNMF) [8] and the proposed method (MGCRDNMF) using the speech data excerpted from the ATR503 database [15] and 5 types of measured noise, namely BusTerminal-5dB, Square-5dB, BowlingAlley-5dB, SubwayStation0dB and DepartmentStore0dB, excerpted from the ATR ambient noise sound database. The test data were created by adding noise signals to clean speech signals with the signal-to-noise ratios (SNRs) of -5, 0 dB. All the audio signals were monaural and sampled at 16KHz. The STFT was computed using the Hanning window with 32ms long and 16ms overlap.

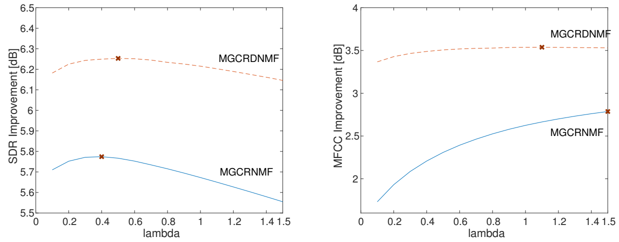

Fig. 2. Average SDR improvement [dB] (left) and MFCC distance improvement [dB] (right) obtained by MGCRNMF and the proposed method using 50 test samples with λ from 0.1 to 1.5 at 0.1 intervals. The points draw the maximum of the curves.

At training phase, 200 utterances spoken by 2 male and 2 female speakers were used to train 40 speech basis spectra. For noise we used the same number of basis spectra. We run 200 iterations for SNMF basis training and 25 iterations for DNMF training. For DNMF, the basis obtained by running NMF for 100 iterations was used as an initialization. The same training set was also used for K-means training. The cluster number was set at 1000. We used 20 order MGCCs with hyperparameter $(\gamma, \alpha) = (-1, 0.42)$ since it has been shown in [8] that this hyperparameters setting can achieve relatively stable high performance under all conditions.

At test time, 50 utterances selected randomly from ATR503 database were used as a test data set. All methods were run 200 iterations. Since each frame should finally converge to one cluster, we do not need to select the closest prototype spectrum at every iteration. Fig. 4 shows that there was only about 1 $\%$ frames would change the cluster to which one it assigned after updating the indicator variables $\{r_t\}$ once so that we can set the updating number of the indicator variables at 1 with considering the balance of the time cost and the performance. Tab. 4 shows a comparison of the runtime between updating indicator variable every iteration and only updating it once using MGCRNMF and the proposed method. The programs were run in the MATLAB 2015b with Inter(R) Xeon E3-1505M V5 CPU @2.80GHz 64bit and 16.0 GB memory. The results show a significant improvement in runtime between w/o update and w/ update. The performance degradation caused by decreasing the updating number was quite small which was only about 0.05 dB. It is noteworthy that without updating every iteration, MGCRNMF can realize a real-time running.

Table 2. From top to bottom, there are respectively average SDR, SIR, MFCC distance Improvement [dB] evaluated under 5 noise conditions. The highest score of each term is shown in bold font type.

Noise Type	SNMF	MGCRNMF	DNMF	Proposed
BusTerminal	10.71	11.22	11.57	**11.94**
Square	6.19	6.45	6.76	**6.88**
BowlingAlley	3.37	3.40	4.01	**4.30**
SubwayStation	3.90	3.78	4.22	**4.46**
DepartmentStore	4.73	4.95	4.76	**4.97**

Noise Type	SNMF	MGCRNMF	DNMF	Proposed
BusTerminal	13.85	15.07	17.43	**18.32**
Square	8.61	9.29	10.32	**11.14**
BowlingAlley	5.27	5.77	6.72	**7.40**
SubwayStation	6.74	7.71	8.27	**8.87**
DepartmentStore	7.09	7.89	8.92	**10.42**

Noise Type	SNMF	MGCRNMF	DNMF	Proposed
BusTerminal	1.79	2.27	3.24	**3.66**
Square	1.87	2.10	1.84	**3.05**
BowlingAlley	1.32	1.97	2.81	**3.13**
SubwayStation	1.81	2.52	2.85	**3.51**
DepartmentStore	1.79	2.27	2.36	**3.22**

We investigated the weight parameter λ during 0.1 to 1.5 at 0.1 intervals and the results are shown in Fig. 4. According to the Fig. 4, we set $\lambda = 0.4$ for MGCRNMF and $\lambda = 0.5$ for the proposed method. We used Signal-to-distortion ratios (SDRs), signal-to-interference ratios (SIRs) [16] and MFCC distance for the evaluation. Given two D-dimension MFCC sequences $x[d]$ and $y[d]$ calculated from N frequency bins, the MFCC distance is defined as follow:

$$Dist = \frac{20D}{N \ln 10} \sqrt{2 \sum_{d}^{D} (x[d] - y[d])^2}. \tag{29}$$

Tab. 4 shows the results of average SDR, SIR and MFCC distance improvement [dB] obtained using SNMF, DNMF, MGCRNMF and the proposed method under 5 noise conditions. The proposed method outperformed the other methods under all the conditions in terms of all the evaluation criteria.

5. CONCLUSION

This paper proposed an unified approach combining a DNMF basis training phase and mel-generalized cepstral regularization NMF which considers an optimization problem combining a NMF-based model-fitting criterion and a MGC regularization which measures IS divergence between the pretrained prototypes and the enhanced speech spectra obtained using a Wiener filter directly at test phase. We derived a computationally efficient algorithm based on majorization-minimization principle. The experimental results showed that the proposed method outperformed the existing methods SNMF, DNMF and MGCRNMF in terms of SDR, SIR and MFCC distance improvements, which showed the effectiveness of the proposed method.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in Adv. NIPS, pp. 556–562, 2000.

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

[3] Y. Xu, J. Du, L. R. Dai and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 23, No. 1, pp. 7–19, 2015.

[4] J. R. Hershey, C. Zhuo, J. L. Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in Proc. ICASSP, pp. 31–35, 2015.

[5] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation.", In Proc. INTERSPEECH, pp. 865–869, 2014.

[6] L. Li, H. Kameoka, and S. Makino, "Discriminative non-negative matrix factorization with majorization-minimization," in Proc. HSCMA, pp. 141–145, 2017.

[7] L. Li, H. Kameoka, T. Higuchi and H. Saruwatari, "Semi-Supervised Joint Enhancement of Spectral and Cepstral Sequences of Noisy Speech," in Proc. INTERSPEECH, pp. 3753–3757, 2016.

[8] L. Li, H. Kameoka, T. Toda and S. Makino, "Speech enhancement using non-negative spectrogram models with mel-generalized cepstral regularization," in Proc. INTERSPEECH, pp. 1998–2002, 2017.

[9] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in ICSLP, Vol. 94, pp. 18–22, 1994.

[10] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in Proc. the 6th International Congress on Acoustics, pp. 17–20, 1968.

[11] T. Masuko, "HMM-based speech synthesis and its applications," Institute of Technology, 2002.

[12] J. D. Leeuw, and W.J. Heiser, "Convergence of correction matrix algorithms for multidimensional scaling," in Geometric representations of relational data, pp. 735–752, 1977.

[13] D. R. Hunter, and K. Lange, "A tutorial on MM algorithms," The American Statistician, 58 (1), pp. 30–37, 2004.

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

[15] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, pp. 357–363, 1990.

[16] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation.", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 4, pp. 1462–1469, 2006.