2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

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

Non-negative matrix factorization (NMF) is effective in terms of signal recovery accuracy for single-channel speech enhancement task, while it does not directly lead to an enhancement in feature domain. To overcome this problem, we have previously proposed an extension of NMF which combines a NMF-model fitting ceriterion and a divergence measuring the NMF model of speech and the mel-generalized cepstral (MGC) representation of a pretrained propotype spectrum. The model has been shown effective both in increasing the signal recovery accuracy and feature domain enhancement. However one drawback is that the regularization term is formulated based on the NMF model of speech which may cause a degradation of the effect of regularization term since the enhanced speech is obtained using Wiener filtering. This paper proposes a novel formulation for MGC regularization and combines it with Discriminative NMF (DNMF) in order to achieve better speech enhancement performance. The experimental results revealed that the proposed method outperformed the perviously proposed model in terms of both the signal recovery accuracy and feature enhancement.

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

1. INTRODUCTION

Speech enhancement is a technique for recovering the speech signal from an obseved noisy speech signal. Since the presence of noise can significantly degrade the qulity of speech transmission systems and the performance of applications such as speech recognition and speech conversion, many efforts have been devoted to increasing the performance of speech enhancement over recent decades.

For monaural speech enhancement task, non-negative matrix factorization (NMF) [1, 2] is a powerful approach attracted a lot of attention since it has been proposed. Althought in the recent years, deep neural networks based machine learning approaches have showed the incredible capability for various supervised audio signal process tasks including monaural speech enhancement [3, 4], NMF still remains attractive under unsupervised setting or only a limited training data set available.

Given an observed magnitude or power spectrogram of a mixture signal, NMF aims to approximate it as the sum of speech and noise spectrogram models, which are represented as a non-negative linear sum of the pretrained basis spectra scaled by time-varying amplitudes. The underlying speech components can be separated out using the Wiener filter constructed by the estimated power spectrograms of speech and noise. Although NMF is shown to be effective in terms of signal recovery accuracy, one drawback is that NMF does not directly lead to an enhancement in the feature domain (e.g., cepstral domain) or in terms of perceived quality. To overcome this drawback, we have previously proposed a NMF framework using mel-generalized cepstral regularization (MGCRNMF) [5], which combines an NMF-based model fitting criterion with a divergence measure between the estimated NMF model of speech and the mel-generalized cepstral (MGC) representation [6] of a prototype spectrum in a pretrained codebook. In [5], we have shown the effectiveness of the MGC regularization in terms of increasing both signal recovery accuracy and cepstral domain enhancement. However, MGCRNMF considers the regularization based on NMF model while the enhanced speech is finally separated out by Wiener filtering, which may lead to a reduction in the effect of the regularization term. To address this problem, this paper proposes a novel formulation of mel-generalized cepstral regularization which measures a divergence between a prototype spectrum represented by MGC representation and the estimated enhanced speech spectra obtained by Wiener filter directly.

On the other hand, Weninger recently proposed a basis training approach called discriminative NMF (DNMF) [7], which trains the basis spectra in such a way that the output of the Wiener filter becomes as close to the spectrogram of each of the training examples as possible so that the separated signals become optimal at test time. Since the basis spectra trained by DNMF also takes the Wiener filter into account, it makes us to believe that the basis spectra trained by DNMF

This work was supported by xxxx.

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

are more appropriate for the proposed formulation than those trained by conventional NMF way. Furthermore, DNMF has been proved it can provide greater separation capability than conventional NMF [7, 8], which further motivates us to connect DNMF with mel-generatlized cepstral regularization in order to achieve better performance of speech enhancement.

The remaining part of the paper proceeds as follows: we review the model of NMF with MGC regularization for speech enhancement task in sec. 2. In sec. 3, we introduce the new formulation for MGC regularization and derive the update rules of paremeters based on majorization-minimization (MM) principle. In the experimental section (sec. 4), we define the data set, investigate the hyperparameters of the model and compare the proposed method with established methods. We conclude this work in sec. 5.

2. NMF WITH MEL-GENERALIZED CEPSTRAL REGULARIZATION

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, we consider approximating it by the sum of speech and noise components, $X_{\omega,t} = X^s_{\omega,t} + X^n_{\omega,t}$, where $X^s_{\omega,t}$ and $X^n_{\omega,t}$ are represented by the non-negative linear combination of K_s speech basis spectra $W^s_{1,\omega}, \ldots, W^s_{K_s,\omega}$ and K_n noise basis spectra $W^n_{1,\omega}, \ldots, W^n_{K_n,\omega}$:

$$X^s_{\omega,t} = \sum_{k=1}^{K_s} W^s_{k,\omega} H^s_{k,t}, \quad X^n_{\omega,t} = \sum_{k=1}^{K_n} W^n_{k,\omega} H^n_{k,t}. \tag{1}$$

In a supervised setting, a concentrated basis matrix $\boldsymbol{W} = [\boldsymbol{W}^s \ \boldsymbol{W}^n]$ are pretrained using speech and noise training samples respectively. $\boldsymbol{H} = [\boldsymbol{H}^s; \boldsymbol{H}^n]$ are the variables to be estimated at test time. NMF leads to different optimization problems according to the definition of the measure of the difference between $\boldsymbol{Y}$ and $\boldsymbol{X} = (X_{\omega,t})_{\Omega\times T}$. Here we use the generalized Kullback Leibler (KL) divergence

$$\mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{\omega,t}\left(Y_{\omega,t}\log\frac{Y_{\omega,t}}{X_{\omega,t}} - Y_{\omega,t} + X_{\omega,t}\right) \tag{2}$$

as a goodness-of-fit criterion. Once $X^s_{\omega,t}$ and $X^n_{\omega,t}$ are estimated, the enhanced speech can be separated out using the Wiener filter constructed with the estimated power spectrogram of speech and noise

$$\hat{\boldsymbol{X}}^s = \frac{\boldsymbol{W}^s\boldsymbol{H}^s}{\boldsymbol{W}\boldsymbol{H}} \otimes \boldsymbol{Y}, \tag{3}$$

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

2.2. Mel-generalized cepstral regularization

When estimating the speech spectrogram $X^s_{\omega,t}$ in a mixture spectrogram $Y_{\omega,t}$, we would want to ensure that the features of $X^s_{\omega,t}$ in cepstral domain are also enhanced. With this motivation, we have proposed a penalty term in [5] defined as the Itakura-Saito (IS) divergence [9] between $X^s_{\omega,t}$ and $S_{\omega,t}(\theta)$

$$\mathcal{J}(\boldsymbol{H}^s, \boldsymbol{\theta}) = \sum_{\omega,t}\Big(\frac{X^s_{\omega,t}}{S_{\omega,t}(\theta)} - \log\frac{X^s_{\omega,t}}{S_{\omega,t}(\theta)} - 1\Big). \tag{4}$$

Here $S_{\omega,t}(\theta) = \beta_{t,r_t}\mu_{\omega,r_t}$ is a scaled pretrained propotype spectrum chosen from I propotypes, which are trained using clean speech samples by k-means. At each iteration, we find the prototype spectrum $\boldsymbol{\mu}_i$ closest to $\boldsymbol{X}^s_t$ in terms of the IS divergence. $r_t \in \{1, \ldots, I\}$ denotes a cluster indicator variable, describing to which of the I clusters the t-th speech spectrum is assigned. To eliminate the scaling indeterminacy, we invoke a scaling parameter β_{t,r_t}, which can be obtained as

$$\hat{\beta}_{t,i} = \frac{1}{\Omega}\sum_{\omega}\frac{X^s_{\omega,t}}{\mu_{\omega,i}}. \tag{5}$$

when $X^s_{\omega,t}$ and $\mu_{\omega,i}$ are given. Thus, the parameter $\boldsymbol{\theta} = \{r_t, \beta_{t,r_t}\}$ consists of a set of cluster indicator variables and corresponding scaling parameters. Note the propotype spectra $\boldsymbol{\mu}_i = [\mu_{1,i}, \ldots, \mu_{\Omega,i}]^\mathsf{T}$ here are represented by MGC representation, which is a parametric model for spectral envelopes of speech described by $M+1$ coefficients and two hyperparameters γ and α:

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

The coefficients $\boldsymbol{c} = [c(0), \ldots, c(M)]^\mathsf{T}$ are called the MGC coefficients (MGCC). 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 α. Here, α must satisfy $|\alpha| < 1$. 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 $\alpha = 0.55$ [10].

2.3. Objective function of MGCRNMF

MGCRNMF considers an optimization problem of minimizing a combined objective function of (2) and (4)

$$\mathcal{F}(\boldsymbol{H},\boldsymbol{\theta}) = \mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{X}) + \lambda\mathcal{J}(\mathbf{H}^s,\boldsymbol{\theta}), \tag{9}$$

where $\lambda \geq 0$ weighs the importance of the mel-generalized cepstral regularization term relative to the NMF cost.

However, since the enhanced speech finally separated out using a Wiener filter, the formulation of MGC regularization based on NMF model of speech $\boldsymbol{W}^s\boldsymbol{H}^s$ may cause a reduction of the effect of the regularization term during the filtering process.

3. PROPOSED METHOD

3.1. A new formulation of MGC regularization

To address the problem mentioned above, we introduce a new formulation of MGC regularization which measures IS divergence between the pretrained propotypes and the enhanced speech spectra obtained using a Wiener filter directly.

$$\tilde{\mathcal{J}}(\boldsymbol{H}^s,\boldsymbol{\theta}) = \sum_{\omega,t}\Big(\frac{\hat{X}^s_{\omega,t}}{S_{\omega,t}(\theta)} - \log\frac{\hat{X}^s_{\omega,t}}{S_{\omega,t}(\theta)} - 1\Big), \tag{10}$$

where, $\hat{X}^s_{\omega,t}$ are the enhanced speech spectra obtained by Wiener filtering (3). With the new regularization term, we can easily obtain the objective function of the proposed method

$$\tilde{\mathcal{F}}(\boldsymbol{H},\boldsymbol{\theta}) = \mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{X}) + \lambda\tilde{\mathcal{J}}(\mathbf{H}^s,\boldsymbol{\theta}). \tag{11}$$

Similarily, $\lambda \geq 0$ here is a weight parameter to measure the importance of the regularization term relative to the NMF cost.

3.2. Relation to DNMF

Instead of applying NMF to speech and noise training samples respectively to train the basis spectra, Weninger [7] proposed directly using the reconstruction error of the separated signals as an objective function for the basis training

$$\begin{aligned} &\text{minimize} \quad f(\boldsymbol{W},\boldsymbol{H}) = \mathcal{D}_{KL}\left(\boldsymbol{T}\,\middle|\,\frac{\boldsymbol{W}^s\boldsymbol{H}^s}{\boldsymbol{W}\boldsymbol{H}}\otimes\boldsymbol{Y}\right) \\ &\text{subject to} \quad \forall k,\ \sum_{\omega} W_{\omega,k} = 1, \end{aligned} \tag{12}$$

where $\boldsymbol{T}$ denotes the spectrograms of clean speech training samples. This framework is called discriminative NMF (DNMF) by analogy with the discriminative models for classification or regression. DNMF trains basis spectra which are optimal to construct a Wiener filter instead of basis spectra which represents spectrograms of clean train samples well, which is more appropriate for the proposed formulation of MGC regularization.

3.3. Update rules and algorithm

Although minimizing the objective function including the regularization term (10) directly is analytically difficult, we can derive a computationally efficient algorithm to find a locally optimal solution based on majorization-minimization (MM) principle [11, 12].

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, 13].

Here, we derive a majorizer for the objective function (11) 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, $\mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{X})$ 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 $\mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{X})$ having a "sum-of-logs" form in the same way as [1]

$$\mathcal{D}_{KL}(\boldsymbol{Y}|\boldsymbol{X}) \leq \mathcal{D}^+_{KL}(\boldsymbol{Y}|\boldsymbol{X}) \tag{13}$$

$$\mathcal{D}^+_{KL}(\boldsymbol{Y}|\boldsymbol{X}) \overset{c}{=} \sum_{\omega,t}\left(-Y_{\omega,t}\sum_k \zeta_{k,\omega,t}\log\frac{W_{k,\omega}H_{k,t}}{\zeta_{k,\omega,t}} + X_{\omega,t}\right),$$

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

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

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 [8]

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

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

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

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:

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

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

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

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

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

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:

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

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

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

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

$$\tilde{\mathcal{F}}^+(\boldsymbol{H}, \boldsymbol{\Gamma}; \hat{\boldsymbol{\theta}}) = \mathcal{D}^+_{KL}(\boldsymbol{Y}|\boldsymbol{X}) + \lambda \tilde{\mathcal{J}}^+(\boldsymbol{H}^s, \boldsymbol{\Gamma}; \hat{\boldsymbol{\theta}}). \tag{27}$$

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:

$$\begin{aligned}
&\sum_{\omega} \frac{\tau_{\omega,t} Y_{\omega,t}}{2\alpha_{k,\omega,t} S_{\omega,t}(\hat{\theta})} {W^s_{k,\omega}}^2 {H^s_{k,t}}^4 + \sum_{\omega} \frac{W^s_{k,\omega}}{\eta_{\omega,t}} {H^s_{k,t}}^3 \\
&\quad - \sum_{\omega} \gamma_{k,\omega,t} {H^s_{k,t}}^2 - \sum_{\omega} \frac{Y_{\omega,t} \xi^3_{k,\omega,t}}{2\tau_{\omega,t} S_{\omega,t}(\hat{\theta}) {W^s_{k,\omega}}^2} = 0, \quad (28) \\
&\sum_{\omega} \frac{W^n_{k,\omega}}{\eta_{\omega,t}} {H^n_{k,t}}^3 - \sum_{\omega} \frac{Y_{\omega,t} \xi^3_{k,\omega,t}}{2\tau_{\omega,t} S_{\omega,t}(\hat{\theta}) {W^n_{k,\omega}}^2} = 0. \quad (29)
\end{aligned}$$

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. 3.3

Require: pretrained speech basis $\boldsymbol{W}$ and I MGC prototypes $\boldsymbol{\mu}$, parameters λ and $MaxIter$

```
random initialize H^s and H^n
for iter = 1 to MaxIter do
    if iter ≤ 50 then
        update H^s and H^n using SSNMF
    else
        calculate the enhanced speech X̂^s using (3)
        for Frame t = 1 to T do
            compute β̂_{t,i} using (5)
            r̂_t = arg min_{r_t} J̃(H^s, θ)
            S_{ω,t}(θ̂) = β̂_{t,r̂_t} μ_{ω,r̂_t}
        end for
        update auxiliary variables Γ using (14), (17), (19),
        (21) , (24) and (26)
        update H^s, H^n by solving the equations (28) and (29)
    end if
end for
```

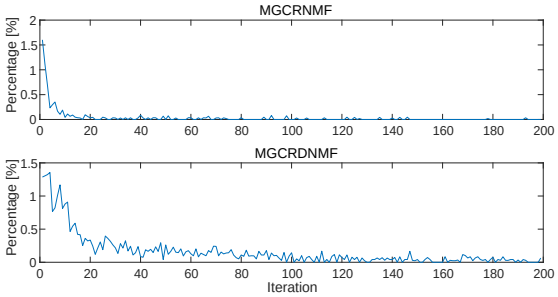

Fig. 1. Percentage of the number of frames which shift the propotype during the updating of MGCRNMF (upper) and MGCRDNMF (bottom). The average number of 10 samples randomly selected under 5 noise types are shown in the figure.

4. EXPERIMENTS

4.1. Experimental conditions

To evaluate the effect of the proposed method for speech enhancement task, we tested supervised NMF (SNMF) [2], Discriminative NMF (DNMF) [8], NMF with mel-generalized cepstral regularization (MGCRNMF) [5] and the proposed method (MGCRDNMF) using the speech data excerpted from the ATR503 database [14] and 5 types of measured noise, respectively 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.

In the 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 with running 100 iterations NMF for 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 [5] that this hyperparameters setting can achieve relatively high performance stably under all conditions.

In the test phase, 50 uttrances selected randomly from ATR503 database were used as test data set. All methods were run 200 iterations while MGCRNMF and the proposed method run 50 iterations NMF as an initialization. During the update, each frame should converge to one cluster, which means we do not need to figure out the cloest propotype spectrum at every iteration. Fig. 1 shows the percentage of how many frames shift the cluster to another one among all the frames at every iteration. The average frame number of the selected test data was about 324 so that 1% of the whole frames was about 3 frames, which means that there was only about 3 frames have not converged to the clusters after updating once. With this result, in the following experiments, we set the update iteration number of the indicator variables $\{r_t\}$ at 1. Tab. 1 shows a comparsion 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 with a very small degradation of the performance about 0.05 dB. It is worth noting that without updating every iteration, MGCRNMF realized a real time computation.

Table 1. A comparsion of runtime [sec] between updating indicator variables at each iteration and updating them at the first iteration of 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

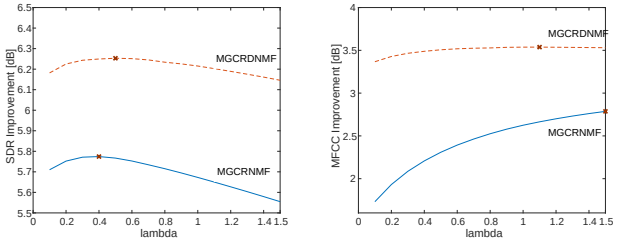

Fig. 2. Average SDR improvement [dB] (left) and MFCC improvement [dB] (right) achieved by MGCRNMF and the proposed method using 50 test samples with λ from 0.1 to 1.5 at 0.1 intervals. The points draw the maximum of the curves.

We investigated the weight parameter λ during 0.1 to 1.5 at 0.1 intervals and the results are shown in Fig. 2. According to the Fig. 2, we set $\lambda = 0.4$ for MGCRNMF and $\lambda = 0.5$ for the proposed method this time in order to achieve the highest signal-to-distortion ratio improvement.

4.2. Objective evaluation

We used Signal-to-distortion ratios (SDRs), signal-to-interference ratios (SIRs) [15] 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:

$$Dist = \frac{20D}{N \ln 10} \sqrt{2 \sum_{d}^{D} (x[d] - y[d])^2}. \qquad (30)$$

Tab. 2 shows the results of average SDR, SIR and MFCC distance improvement [dB] obtained using SNMF, DNMF,

Table 2. From top to bottom, there are respectively average SDR, SIR, MFCC 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**

MGCRNMF and the proposed method under 5 noise conditions. The proposed method outperformed the other methods under all the conditions in all the evaluation criteria.

5. CONCLUSION

This paper proposed a novel formulation for mel-generalized cepstral regularization to enhance speech in spectral and cepstral domain, which takes the Wiener filtering process into account. We combined the proposed regularization with Discriminative NMF basis training approach and derived a computationally efficient algorithm based on majorization-minimization principle. The experimental results showed that the proposed algorithm outperformed SNMF, DNMF and previously proposed MGCRNMF in 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] L. Li, H. Kameoka, T. Toda and S. Makino, "Speech enhancement using non-negative spectrogram models with mel-generalized cepstral regularization," in Proc. INTERSPEECH, 2017.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

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

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

[14] 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.

[15] 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, 2016.