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

Speech enhancement using non-negative spectrogram models with mel-generalized cepstral regularization

Li Li[1], *Hirokazu Kameoka*[2], *Tomoki Toda*[3] *and Shoji Makino*[1]

[1]University of Tsukuba, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[3]Information Technology Center, Nagoya University, Japan

lili@mmlab.cs.tsukuba.ac.jp, kameoka.hirokazu@lab.ntt.co.jp
tomoki@icts.nagoya-u.ac.jp, maki@tara.tsukuba.ac.jp

Abstract

Spectral domain speech enhancement algorithms based on non-negative spectrogram models such as non-negative matrix factorization (NMF) and non-negative matrix factor deconvolution are powerful in terms of signal recovery accuracy, however they do not directly lead to an enhancement in the feature domain (e.g., cepstral domain) or in terms of perceived quality. We have previously proposed a method that makes it possible to enhance speech in the spectral and cepstral domains simultaneously. Although this method was shown to be effective, the devised algorithm was computationally demanding. This paper proposes yet another formulation that allows for a fast implementation by replacing the regularization term with a divergence measure between the NMF model and the mel-generalized cepstral (MGC) representation of the target spectrum. Since the MGC is an auditory-motivated representation of an audio signal widely used in parametric speech synthesis, we also expect the proposed method to have an effect in enhancing the perceived quality. Experimental results revealed the effectiveness of the proposed method in terms of both the signal-to-distortion ratio and the cepstral distance.

Index Terms: speech enhancement, non-negative matrix factorization, mel-generalized cepstral representation, single channel signal processing

1. Introduction

Enhancing speech signals from observed noisy speech signals is an important task in various applications since the presence of noise can significantly degrade the quality of speech transmission systems and the performance of e.g., speech recognition and speech conversion.

In recent years, many spectral domain speech enhancement algorithms based on non-negative spectrogram models, such as non-negative matrix factorization (NMF) [1,2] and non-negative factor deconvolution [3], have been proposed. These methods are particularly noteworthy in that they can work successfully even without any prior knowledge about the noise characteristics by using a reasonable assumption that speech and noise are additive in the spectral domain. In a semi-supervised setting, the aim of these methods is to approximate the magnitude (or power) spectrogram of an observed mixture signal as the sum of speech and noise spectrogram models using the pretrained spectral templates of speech. The underlying speech components can be separated out using the Wiener filter constructed by the estimated power spectrograms of speech and noise. Although these methods are shown to be effective in terms of signal recovery accuracy, they do not directly lead to an enhancement in the feature domain (e.g., cepstral domain) or in terms of perceived quality. Hence, naively using these algorithms as a front-end processing for e.g., speech recognition and voice conversion does not always lead to satisfactory results. To overcome this drawback, we have previously proposed a semi-supervised NMF framework using cepstral distance regularization (CDRNMF) [4], which makes it possible to enhance speech in both the spectral and cepstral domains by optimizing a combined objective function 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 coefficients (MFCC) domain. The objective function that contains the explicit form of the mapping function from the spectral domain to the MFCC domain becomes very difficult to optimize, resulting in an optimization algorithm with slow convergence.

This paper proposes yet another formulation with an objective function that is relatively easy to optimize. Specifically, we combine an NMF-based model fitting criterion with a divergence measure between the NMF model and the mel-generalized cepstral (MGC) representation [5] of a prototype spectrum in a pretrained codebook. An MGC representation is a parametric model for a spectral envelope described by a set of coefficients called the MGC coefficients, which takes the all-pole spectral model and the cepstral representation as special cases. Through the hyperparameter setting, the MGC representation can be set to have a mel-scale frequency resolution similar to the human auditory system. This hyperparameter setting makes the MGC representation an auditory-motivated representation of audio signals and is widely used in parametric speech synthesis. This motivates us to believe that the proposed method will also have an effect in enhancing the preceived quality of the generated speech.

2. Formulation

2.1. Speech enhancement with NMF

We start by reviewing a speech enhancement method using NMF. 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 $W^{(s)}_{1,\omega}, \dots, W^{(s)}_{K_s,\omega}$ and K_n noise basis spectra $W^{(n)}_{1,\omega}, \dots, W^{(n)}_{K_n,\omega}$:

$$X^{(s)}_{\omega,t} = \sum_{k=1}^{K_s} W^{(s)}_{k,\omega} H^{(s)}_{k,t}, \;\; X^{(n)}_{\omega,t} = \sum_{k=1}^{K_n} W^{(n)}_{k,\omega} H^{(n)}_{k,t}. \qquad (1)$$

In a semi-supervised setting, the speech basis spectra are pretrained using clean speech training samples. $\boldsymbol{W}^{(n)}$, $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(n)}$ are the variables to be estimated at test time. Here, we

http://dx.doi.org/10.21437/Interspeech.2017-1492

Table 1: *Examples of α and γ settings under 16 kHz sampling frequency.*

	$\gamma = -1$	$\gamma = 0$
$\alpha = 0$	All-pole model	Cepstral representation
$\alpha = 0.42$	Warped all-pole model	Mel-cepstral representation

use the generalized Kullback Leibler (KL) divergence

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

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.

As stated above, this method does not directly lead to an enhancement in the cepstral domain, implying that naively using it as a front-end for e.g., speech recognition and voice conversion does not always lead to satisfactory results. To address this drawback, we introduce a novel regularization method using the MGC representation of a prototype spectrum in a pretrained codebook with the aim of improving the quality of the enhanced speech in the cepstral domain.

2.2. Mel-generalized cepstral representation

The mel-generalized cepstral (MGC) representation $S(\omega)$ is a parametric model for spectral envelopes of speech described by $M+1$ coefficients and two hyperparameters γ and α:

$$\begin{aligned} S(\omega) &= l_\gamma^{-1}\left(\sum_{m=0}^{M} c(m)\Psi_\alpha^m(e^{j\omega}) \right) \quad (3) \\ &= \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c(m)\Psi_\alpha^m(e^{j\omega})\right)^{1/\gamma} & (0 < |\gamma| \le 1) \\ \exp \sum_{m=0}^{M} c(m)\Psi_\alpha^m(e^{j\omega}) & (\gamma = 0) \end{cases}, \end{aligned}$$

where the coefficients $\boldsymbol{c} = [c(0), \dots, c(M)]^\mathsf{T}$ are called the MGC coefficients (MGCC). The function $l_\gamma^{-1}(\cdot)$ is the inverse of the generalized logarithmic function

$$l_\gamma(\omega) = \begin{cases} (\omega^\gamma - 1)/\gamma & (0 < |\gamma| \le 1) \\ \log \omega & (\gamma = 0) \end{cases}, \quad (4)$$

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

$$\Psi_\alpha(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad (5)$$

which can be seen as a frequency warping function parameterized by α. Here, α must satisfy $|\alpha| < 1$. We can confirm that $S(\omega)$ 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 $\alpha = 0.55$ [6]. Tab. 1 shows the four (γ, α) settings we investigated in our experiments.

2.3. Objective function

When estimating the speech spectrogram $X^{(s)}_{\omega,t}$ in a mixture spectrogram $Y_{\omega,t}$, we would want to ensure that the MGCCs of $X^{(s)}_{\omega,t}$ are also enhanced. Here, we introduce a penalty term defined as the Itakura-Saito (IS) divergence [7] between $X^{(s)}_{\omega,t}$ and $S_{\omega,t}(\theta)$

$$\mathcal{J}(\boldsymbol{W}^{(s)}, \boldsymbol{H}^{(s)}, \boldsymbol{\theta}) = \sum_{\omega,t} \left\{ \frac{X^{(s)}_{\omega,t}}{S_{\omega,t}(\theta)} - \log \frac{X^{(s)}_{\omega,t}}{S_{\omega,t}(\theta)} - 1 \right\}, \quad (6)$$

where $S_{\omega,t}(\theta)$ is the target MGC representation at frame t and frequency ω defined using the expression (3) described by parameter θ. In sec. 3, we introduce two ways to model $S_{\omega,t}(\theta)$, which leads to different θ representations and algorithms.

The proposed method considers an optimization problem of minimizing a combined objective function of (2) and (6)

$$\mathcal{F}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{Y}|\boldsymbol{X}) + \lambda \mathcal{J}(\mathbf{W}^{(s)}, \mathbf{H}^{(s)}, \boldsymbol{\theta}), \quad (7)$$

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

3. Algorithms

In this section, we derive optimization algorithms for $\boldsymbol{W}^{(n)}$, $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(n)}$ based on the majorization-minimization (MM) principle [8, 9] when target MGC representation $S_{\omega,t}(\hat{\theta})$ with θ fixed to $\hat{\theta}$. The first algorithm considers a frame-independent representation of $S_{\omega,t}(\theta)$ whereas the the second algorithm considers $S_{\omega,t}(\theta)$ as a prototype spectrum drawn from a codebook of MGCCs pretrained using the K-means clustering.

3.1. Update rules based on majorization-minimization

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, 10].

Here, we derive a majorizer for the objective function (7) with respect to $\boldsymbol{W}$ and $\boldsymbol{H}$. First, $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X})$ 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 $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X})$ having a "sum-of-logs" form in the same way as [1]

$$\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X}) \le \mathcal{I}^+(\boldsymbol{Y}|\boldsymbol{X}) \quad (8)$$

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

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 (8) holds if and only if

$$\zeta_{k,\omega,t} = \frac{W_{k,\omega}H_{k,t}}{\sum_{k'} W_{k',\omega}H_{k',t}}. \quad (9)$$

Next, we focus on the regularization term (6). We can see that the second term in (6) also has a "log-of-sum" form and thus we can employ Jensen's inequality

$$\mathcal{J}(\boldsymbol{W}^{(s)}, \boldsymbol{H}^{(s)}; \hat{\boldsymbol{\theta}}) \le \mathcal{J}^+(\boldsymbol{W}^{(s)}, \boldsymbol{H}^{(s)}, \boldsymbol{\eta}; \hat{\boldsymbol{\theta}}) \quad (10)$$

Algorithm 1 Algorithm presented in subsec. 3.2

```
Require: pretrained speech basis $\boldsymbol{W}^{(s)}$, parameters $\lambda$, $\alpha$, $\gamma$ and $M$, $MaxIter$
randomly initialize $\boldsymbol{W}^{(n)}$, $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(n)}$
for $iter = 1$ to $MaxIter$ do
    if $iter \leq 50$ then
        update $\boldsymbol{W}^{(n)}$, $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(n)}$ using SSNMF
    else
        $\boldsymbol{X}^{(s)} = \boldsymbol{W}^{(s)}\boldsymbol{H}^{(s)}$
        $\boldsymbol{c} = MGC(\boldsymbol{X}^{(s)}, M, \alpha, \gamma)$
        compute $S_{\omega,t}(\theta)$ using (17)
        update $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(n)}$ using (15) and (16)
        update $\boldsymbol{W}^{(n)}$ using (14)
    end if
end for
```

$$\mathcal{J}^+(\boldsymbol{W}^{(s)}, \boldsymbol{H}^{(s)}, \boldsymbol{\eta}; \hat{\boldsymbol{\theta}}) \overset{c}{\leq} \sum_{\omega,t}\left\{\frac{X^{(s)}_{\omega,t}}{S_{\omega,t}(\hat{\theta})} - \sum_{k}^{K_s} \eta_{k.\omega,t} \log \frac{W^{(s)}_{k,\omega}H^{(s)}_{k,t}}{\eta_{k,\omega,t}}\right\},$$

where $\eta_{k,\omega,t} > 0$ is a weight parameter that sums to unity, $\sum_k \eta_{k,\omega,t} = 1$. It can be shown that equality of this inequality holds if and only if

$$\eta_{k,\omega,t} = \frac{W^{(s)}_{k,\omega}H^{(s)}_{k,t}}{\sum_{k'} W^{(s)}_{k',\omega}H^{(s)}_{k',t}}. \qquad (11)$$

The upper bound for the objective function can be easily obtained by combining the majorizers for each term as

$$\mathcal{F}^+(\boldsymbol{W}, \boldsymbol{H}; \hat{\boldsymbol{\theta}}) = \mathcal{I}^+(\boldsymbol{Y}|\boldsymbol{X}) + \lambda\mathcal{J}^+(\boldsymbol{W}^{(s)}, \boldsymbol{H}^{(s)}, \boldsymbol{\eta}; \hat{\boldsymbol{\theta}}). \quad (12)$$

The update rules for $W_{\omega,k}$ and $H_{k,t}$ can be obtained by setting at zeros the partial derivatives of the derived majorizer with respect to $W^{(s)}_{k,\omega}$, $W^{(n)}_{k,\omega}$, $H^{(s)}_{k,t}$ and $H^{(n)}_{k,t}$:

$$W^{(s)}_{\omega,t} \leftarrow W^{(s)}_{\omega,t}\frac{\sum_t Y_{\omega,t}H^{(s)}_{k,t}/X_{\omega,t} + \lambda\sum_t H^{(s)}_{k,t}/X^{(s)}_{\omega,t}}{\sum_t H^{(s)}_{k,t} + \lambda\sum_t H^{(s)}_{k,t}/S_{\omega,t}(\hat{\theta})}, \qquad (13)$$

$$W^{(n)}_{\omega,t} \leftarrow W^{(n)}_{\omega,t}\frac{\sum_t Y_{\omega,t}H^{(n)}_{k,t}/X_{\omega,t}}{\sum_t H^{(n)}_{k,t}}, \qquad (14)$$

$$H^{(s)}_{\omega,t} \leftarrow H^{(s)}_{k,t}\frac{\sum_\omega Y_{\omega,t}W^{(s)}_{k,\omega}/X_{\omega,t} + \lambda\sum_\omega W^{(s)}_{k,\omega}/X^{(s)}_{\omega,t}}{\sum_\omega W^{(s)}_{k,\omega} + \lambda\sum_\omega W^{(s)}_{k,\omega}/S_{\omega,t}(\hat{\theta})}, \qquad (15)$$

$$H^{(n)}_{\omega,t} \leftarrow H^{(n)}_{k,t}\frac{\sum_\omega Y_{\omega,t}W^{(n)}_{k,\omega}/X_{\omega,t}}{\sum_\omega W^{(n)}_{k,\omega}}. \qquad (16)$$

Since the regularization term only impacts on the estimation of speech components, the update rules for $W^{(n)}_{k,\omega}$ and $H^{(n)}_{k,t}$ are the same as the regular NMF algorithm [1].

3.2. Regularized algorithm using frame-independent MGC representation

In the rest of sec. 3, we introduce two approaches to modeling the target MGC representation $\boldsymbol{S}$.

Algorithm 1 shows an algorithm using the frame-independent MGC regularization. Here, $S_{\omega,t}(\theta)$ is defined as

$$S_{\omega,t}(\theta) = l_\gamma^{-1}\left(\sum_{m=0}^{M} c_{m,t}\Psi^m_\alpha(e^{j\omega})\right). \qquad (17)$$

Algorithm 2 Algorithm presented in subsec. 3.3

```
Require: pretrained speech basis $\boldsymbol{W}^{(s)}$ and $I$ MGC prototypes $\boldsymbol{\mu}$, parameters $\lambda$ and $MaxIter$
random initialize $\boldsymbol{W}^{(n)}$, $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(n)}$
for $iter = 1$ to $MaxIter$ do
    if $iter \leq 50$ then
        update $\boldsymbol{W}^{(n)}$, $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(n)}$ using SSNMF
    else
        $\boldsymbol{X}^{(s)} = \boldsymbol{W}^{(s)}\boldsymbol{H}^{(s)}$
        for Frame $t = 1$ to $T$ do
            compute $\hat{\beta}_{t,i}$ using (19)
            $\hat{r}_t = \arg\min_{r_t} \mathcal{J}(\boldsymbol{W}^{(s)}, \boldsymbol{H}^{(s)}, \boldsymbol{\theta})$
            $S_{\omega,t}(\hat{\theta}) = \hat{\beta}_{t,\hat{r}_t}\mu_{\omega,\hat{r}_t}$
        end for
        update $\boldsymbol{H}^{(s)}$ and $\boldsymbol{H}^{(n)}$ using (15) and (16)
        update $\boldsymbol{W}^{(n)}$ using (14)
    end if
end for
```

Here, the parameter $\boldsymbol{\theta} = \boldsymbol{c} = \{c_{m,t}\}$ is a set of MGCCs, which must be estimated in a data-driven manner. This regularization aims to ensure the smoothness of the spectral envelopes of $X^{(s)}_{\omega,t}$.

3.3. Regularized algorithm using I MGC prototypes

Here, we introduce another algorithm, which utilizes I MGC prototypes of clean speech pretrained using the K-means algorithm. We use the following objective function for the prototype training

$$\mathcal{C}_{\mathrm{IS}}(\boldsymbol{\mu}, \boldsymbol{r}) = \sum_{\omega,t}\left\{\frac{Y^{(s)}_{\omega,t}}{\mu_{\omega,r_t}} - \log\frac{Y^{(s)}_{\omega,t}}{\mu_{\omega,r_t}} - 1\right\}, \qquad (18)$$

where $\mu_{\omega,i}$ denotes the prototype spectrum associated with cluster i, $Y^{(s)}_{\omega,t}$ denotes the power spectrum of a clean speech sample and $r_t \in \{1, \ldots, I\}$ denotes a cluster indicator variable, describing to which of the I clusters the t-th speech spectrum is assigned. Note that this objective differs from the regular K-means algorithm in that the distance between a data sample and a cluster prototype is measured using the IS divergence.

At test time, we find the prototype spectrum $\boldsymbol{\mu}_i$ closest to $\boldsymbol{X}^{(s)}_t$ in terms of the IS divergence at each iteration where $S_{\omega,t}(\theta) = \beta_{t,r_t}\mu_{\omega,r_t}$. Here, the parameter $\boldsymbol{\theta} = \{r_t, \beta_{t,r_t}\}$ consists of a set of cluster indicator variables and corresponding scaling parameters introduced to eliminate the scaling indeterminacy. When $X^{(s)}_{\omega,t}$ and $\mu_{\omega,i}$ are given, we can easily obtain the optimal scaling as

$$\hat{\beta}_{t,i} = \frac{1}{\Omega}\sum_{\omega}\frac{X^{(s)}_{\omega,t}}{\mu_{\omega,i}}. \qquad (19)$$

4. Experiments

To evaluate the effect of the proposed method for speech enhancement task, we tested the baseline semi-supervised NMF (SSNMF) [2], and the proposed algorithm 1 and 2 using the speech data excerpted from the ATR503 database [11] and two types of measured noise, namely departement store and subway station, excerpted from the ATR ambient noise sound database. We used Signal-to-distortion ratios (SDRs), signal-to-interference ratios (SIRs) [12] and MFCC distance for the

Table 2: *SDR, SIR and MFCC distance improvement [dB] obtained by the baseline SSNMF and the proposed algorithms under noise measured in a department store and a subway station. The highest score of each term is shown in bold type.*

			Input SNR = -5 dB			Input SNR = 0 dB		
			Δ SDR	Δ SIR	Δ MFCC	Δ SDR	Δ SIR	Δ MFCC
SSNMF (baseline)			4.05	5.23	1.63	3.64	5.32	1.30
Algorithm 1	Order = 12	(-1,0.42)	**5.21**	7.37	1.87	3.38	6.95	1.44
		(-1,0)	5.10	7.26	2.06	3.28	**8.24**	1.34
		(0,0.42)	4.64	6.29	1.70	3.57	5.39	1.21
		(0,0)	4.96	6.70	1.83	3.28	5.71	1.14
	Order = 20	(-1,0.42)	4.69	7.22	1.99	3.40	6.41	1.54
		(-1,0)	4.98	6.87	1.97	3.24	7.69	1.65
		(0,0.42)	3.83	5.11	1.32	4.05	5.58	1.34
		(0,0)	4.70	6.66	1.61	3.75	5.66	1.20
Algorithm 2	Order = 20	(-1,0.42)	5.00	8.01	**2.54**	**4.25**	8.02	**1.86**
		(-1,0)	4.41	**8.32**	2.41	3.86	7.11	1.80
		(0,0.42)	4.70	7.83	2.27	4.09	7.30	1.78
		(0,0)	4.59	8.02	1.86	4.12	6.77	1.62

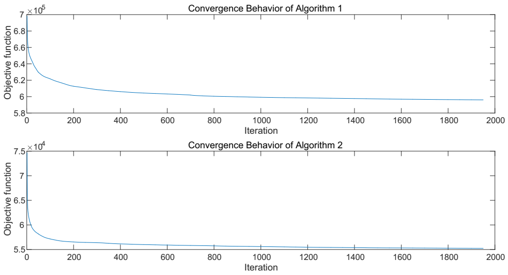

Figure 1: *Convergence behavior of the proposed algorithms. Algorithm 1 is shown in the top figure and algorithm 2 is shown in the bottom one.*

Table 3: *Computational time of SSNMF, CDRNMF and the proposed methods.*

	rumtime/iter [sec]	iteratin number
SSNMF [2]	0.0297	1000
CDRNMF [4]	0.4317	500000
Algorithm 1	0.8181	2000
Algorithm 2	3.4218	2000

evaluation. Given two D-dimension MFCC sequences $x[d]$ and $y[d]$ calculated from N frequency bins, the MFCC distance is defined as follow:

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

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 50 speech basis spectra. For noise we used the same number of basis spectra. The same train set was also used for K-means training. The cluster number was set at 1000. We investigated MGCC with 12 order and 20 order repectively with four special cases shown as hyperparameters $(\gamma, \alpha) = \{(-1, 0.42), (-1, 0), (0, 0.42), (0, 0)\}$. We run 50 iterations SSNMF at the beginning of every algorithm as the initialization.

First, we show the convergence behavior of the proposed algorithms. For comparsion, we also show the runtime of every iteration and iteration demanding of SSNMF and CDRNMF in Tab. 3. All the programs were run in the MATLAB 2015b with Inter(R) Core(TM) i5-5200U CPU @2.20GHz 64bit and 8.0 GB memory. Fig. 1 shows the proposed algorithms converged quickly. Comparing with CDRNMF, though the time of every iteration became longer, the processing time decreased significantly since the proposed algorithms converged with fewer iterations.

Tab. 2 shows the results of SDR, SIR and MFCC distance improvement [dB] obtained by SSNMF and the proposed algorithms under two types of noise. Both of the proposed algorithms achieved an improvement in all the evaluation criteria with appropriate hyperparameters $(\lambda, \gamma, \alpha)$, which showed the effectiveness of the proposed algorithms in jointly enhancing speech in spectral and cepstral domain. As shown in the Tab. 2, the algorithm 2 achieved higher SDR, SIR and MFCC distance improvement on average than algorithm 1. Although $(\gamma, \alpha) = (-1, 0.42)$ did not achieve the highest score under all the conditions, it still can be thought as the first choice of hyperparameters since the stationary improvement shown in the table.

5. Conclusions

This paper proposed a novel approach to jointly enhance speech in spectral and cepstral domain, which allows for a fast implementation. The method optimizes a combined objective function of an NMF-based model fitting criterion with a divergence measure between the NMF model and the mel-generalized cepstral representation in the spectral domain. We derived two algorithms based on majorization-minimization principle. The experimental results showed that both of the proposed algorithms outperformed the baseline SSNMF in SDR, SIR and MFCC distance improvement, which showed the effectiveness of mel-generalized cepstral regularization.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 26730100 and 17H01763, and SECOM Science and Technology Foundation.

7. 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] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in International Conference on Independent Component Analysis and Signal Separation, pp. 494–499, Springer Berlin Heidelberg, 2004.

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

[5] 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.

[6] T. Masuko, “HMM-based speech synthesis and its applications,” Institute of Technology, 2002.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.