INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Semi-Supervised Joint Enhancement of Spectral and Cepstral Sequences of Noisy Speech

Li Li[1], *Hirokazu Kameoka*[1,2], *Takuya Higuchi*[2] *and Hiroshi Saruwatari*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
{li_li, hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp
{kameoka.hirokazu, higuchi.takuya}@lab.ntt.co.jp

Abstract

While spectral domain speech enhancement algorithms using non-negative matrix factorization (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. This implies that naively using these algorithms as front-end processing for e.g., speech recognition and speech conversion does not always lead to satisfactory results. To address this problem, this paper proposes a novel method that aims to jointly enhance the spectral and cepstral sequences of noisy speech, by optimizing a combined objective function consisting of an NMF-based model-fitting criterion defined in the spectral domain and a Gaussian mixture model (GMM)-based probability distribution defined in the cepstral domain. We derive a novel majorizer for this objective function, which allows us to derive a convergence-guaranteed iterative algorithm based on a majorization-minimization scheme for the optimization. Experimental results revealed that the proposed method outperformed the conventional NMF approach in terms of both signal-to-distortion ratio and cepstral distance.

Index Terms: speech enhancement, Gaussian mixture model, non-negative matrix factorization, mel-frequency cepstral coefficients, majorization-minimization

1. Introduction

The presence of noise in speech can significantly degrade the quality of speech transmission systems and the performance of speech processing systems including speech recognition and voice conversion. Many speech enhancement algorithms have been proposed with the goal of overcoming this problem. Conventional speech enhancement methods can be roughly divided into two types, according to the domain in which the enhancement is performed, namely feature domain methods and spectral domain methods. The former aims to recover clean speech features (typically cepstral features) whereas the latter aims to recover clean speech spectra (or signals).

Since feature domain methods directly enhance the features of speech, they are particularly useful as a front-end for speech processing systems that use speech features as inputs. Stereo piecewise linear compensation for environment (SPLICE) [1,2] is a typical example of such methods. With this approach, a Gaussian mixture model (GMM) is used to model the joint distribution of clean and noisy speech features. The GMM is trained using stereo synchronous data of noisy and clean speech samples. By using the trained GMM, a mapping function from a noisy speech feature to its clean version is defined as the conditional expectation of a clean speech feature given a noisy observation. Since each of the Gaussians represents a linear transform, the mapping is piecewise linear. Although it has been shown that it yields a steady improvement in speech recognition performance, one drawback is that the performance tends to be poor when the test condition does not match the training condition (when we face unseen noise). Adaptation techniques can be used to compensate for this mismatch [3], but the noise characteristics of the test condition must not differ significantly from those of the training condition for these techniques to work successfully. Other feature domain methods have more or less the same limitations [4–7].

In contrast to the feature domain methods, spectral domain methods are particularly noteworthy in that they can work successfully even without any prior knowledge about the noise characteristics. This is because we can make use of a reasonable spectrum model to estimate the underlying speech components in an observed spectrum thanks to the additive nature of speech and noise components in the spectral domain. The semi-supervised non-negative matrix factorization (NMF) approach [8] is an example of such methods, and it has attracted a lot of attention in recent years. Factorizing the magnitude (or power) spectrogram of a mixture signal, interpreted as a non-negative matrix, into the product of two non-negative matrices can be interpreted as approximating the observed spectra at each time frame as a linear sum of basis spectra scaled by time-varying amplitudes. This amounts to decomposing the observed spectrogram into the sum of low rank spectrograms. In a semi-supervised setting, the basis spectra of speech are firstly trained using clean speech samples. NMF is then applied to an observed noisy speech spectrogram, where a subset of the basis spectra is fixed at the pretrained spectra. In this way, we can separate out the underlying speech components using the Wiener filter obtained with the estimated speech and noise spectrograms. Although this approach is powerful in terms of a signal-to-noise ratio measure or some subjective criteria, one drawback is that they do not necessarily lead to an improvement in the quality of the enhanced speech in the feature domain. This implies that naively using these algorithms as front-end processing for applications such as speech recognition and voice conversion does not always lead to satisfactory results.

As stated above, feature domain methods (e.g., SPLICE) and spectral domain methods (e.g., NMF) have their own advantages and disadvantages. To address the drawbacks and combine the advantages of these methods, this paper proposes a novel approach that aims to jointly enhance the spectral and cepstral sequences of noisy speech, by optimizing a combined objective function consisting of an NMF-based model-fitting criterion defined in the spectral domain and a GMM-based probability distribution defined in the cepstral domain. We derive a novel majorizer for this objective function, which allows us to derive a convergence-guaranteed iterative algorithm based on a majorization-minimization scheme for optimization.

2. Formulation

2.1. Problem setting

We start by reviewing the formulation of the NMF approach. Let us denote an observed magnitude (or power) spectrogram as $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$, where ω and t are frequency and time indices. Given an observed spectrogram $\mathbf{Y}$, we consider approximating it with the sum of the 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 non-negative linear combinations of K_s speech basis spectra $H^{(s)}_{1,\omega}, \ldots, H^{(s)}_{K_s,\omega}$ and K_n noise basis

http://dx.doi.org/10.21437/Interspeech.2016-1286

spectra $H^{(n)}_{1,\omega},\ldots,H^{(n)}_{K_n,\omega}$:

$$X^{(s)}_{\omega,t}=\sum_{k=1}^{K_s}H^{(s)}_{k,\omega}U^{(s)}_{k,t},\quad X^{(n)}_{\omega,t}=\sum_{k=1}^{K_n}H^{(n)}_{k,\omega}U^{(n)}_{k,t}. \tag{1}$$

In a semi-supervised setting, the speech basis spectra $H^{(s)}_{1,\omega},\ldots,H^{(s)}_{K_s,\omega}$ are pretrained using clean speech samples. Thus, $\boldsymbol{U}^{(s)}$, $\boldsymbol{H}^{(n)}$ and $\boldsymbol{U}^{(n)}$ are the unknown variables to be estimated in the separation process. 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 I divergence

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

as the goodness-of-fit criterion. Once $\boldsymbol{U}^{(s)}$, $\boldsymbol{H}^{(n)}$ and $\boldsymbol{U}^{(n)}$ are obtained with an NMF algorithm, we can separate out the underlying speech components using the Wiener filter constructed with $\boldsymbol{X}^{(s)}=(X^{(s)}_{\omega,t})_{\Omega\times T}$ and $\boldsymbol{X}^{(n)}=(X^{(n)}_{\omega,t})_{\Omega\times T}$. As stated above, this method does not necessarily lead to an improvement in the quality of the enhanced speech in the feature domain, implying that naively using this method as a front-end for such applications as speech recognition and voice conversion does not always produce satisfactory results. When performing separation, we would want to ensure that the cepstral feature of the separated component is also enhanced. With this as motivation, we introduce a probability density function over the cepstral representation of $\boldsymbol{X}^{(s)}$ to define an additional cost function. Here, we define it with the logarithm of a GMM in the MFCC domain

$$\mathcal{K}(\boldsymbol{X}^{(s)})=\log\prod_t\sum_m w_m\prod_n\mathcal{N}(\mathcal{X}^{(s)}_{n,t};\mu_{n,m},\sigma^2_{n,m}), \tag{3}$$

$$\mathcal{X}^{(s)}_{n,t}=\sum_l c_{n,l}\log\sum_\omega f_{l,\omega}X^{(s)}_{\omega,t}, \tag{4}$$

where $\boldsymbol{\mathcal{X}}^{(s)}_t=(\mathcal{X}^{(s)}_{0,t},\ldots,\mathcal{X}^{(s)}_{N-1,t})^{\mathsf{T}}$ is an N-dimensional mel-frequency cepstral coefficient (MFCC) vector of $X^{(s)}_{0,t},\ldots,X^{(s)}_{\Omega-1,t}$. Here, $f_{l,\omega}$ is the lth coefficient of the mel-filter banks and $\{c_{n,l}\}_{0\le n\le N-1,0\le l\le L-1}$ are the coefficients of the discrete cosine transform. Note that this expression reduces to the log mel-spectrum when $c_{m,n}=\delta_{m,n}$ (where δ denotes Kronecker's delta), and the log-power spectrum when $f_{n,\omega}=\delta_{n,\omega}$ and $c_{m,n}=\delta_{m,n}$, implying that the following derivation does not restrict the feature vector definition to MFCC alone. $\theta=\{\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m,w_m\}_{1\le m\le M}$ is a set consisting of the GMM parameters where $\boldsymbol{\mu}_m=(\mu_{1,m},\ldots,\mu_{N,m})^T$, $\boldsymbol{\Sigma}_m=\mathrm{diag}(\sigma_{1,m},\ldots,\sigma_{N,m})$ and w_m are the mean, covariance and weight of the mth Gaussian components. As with the speech basis spectra, θ is pretrained from clean speech samples. Thus, the greater Eq. (4) becomes, the more likely $\boldsymbol{X}^{(s)}$ is to be enhanced in the feature domain.

The proposed method considers an optimization problem that consists of minimizing a combined objective function of Eqs. (2) and (4)

$$\mathcal{J}(\boldsymbol{U}^{(s)},\boldsymbol{H}^{(n)},\boldsymbol{U}^{(n)})=\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X})-\lambda\mathcal{K}(\boldsymbol{X}^{(s)}), \tag{5}$$

where $\lambda\ge 0$ weighs the importance of the MFCC-GMM term relative to the NMF cost. This optimization problem is mathematically challenging in the sense that the objective function simultaneously involves a spectral distance term $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X})$ and a cepstral distance term $\mathcal{K}(\boldsymbol{X}^{(s)})$. Although $\mathcal{K}(\boldsymbol{X}^{(s)})$ is simply a log-GMM when viewed as a function of cepstral parameters $\{\mathcal{X}^{(s)}\}$, it becomes more complicated when viewed as a function of spectral parameters $\{X^{(s)}\}$. This paper proposes a novel general framework for solving this class of optimization problems.

2.2. Proposed method seen as regularized NMF

The present problem setting can be seen as a regularized variant of NMF if we interpret $-\mathcal{K}(\boldsymbol{X}^{(s)})$ as a regularization term. When some of the speech and noise basis spectra become similar, the decomposition of an observed spectrum into speech and noise components will not be unique. In such a case, components originating from speech can be misinterpreted as noise components and vice versa, leading to an inaccurate separation. The regularization term is expected to play the role of eliminating this kind of indeterminacy by keeping the estimated speech spectra within a proper range in the MFCC domain. In this sense, the regularization term can contribute more to improving the separation performance than the regular (unregularized) NMF. This effect will be confirmed in Sec. 3.

2.3. Majorization-minimization algorithm

Although it is difficult to solve the above optimization problem analytically, we can develop a computationally efficient algorithm to find a locally optimal solution based on a majorization-minimization method. Note that the majorization-minimization method itself is not an algorithm, but a description of how to construct an optimization algorithm. When applying the majorization-minimization method to the problem of minimizing a certain objective function, the first step is to design an auxiliary function called a "majorizer" that never lies below the objective function. Suppose $F(\Theta)$ is an objective function that we wish to minimize with respect to Θ. A majorizer $F^+(\Theta,\alpha)$ is then 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. For example, the expectation-maximization (EM) algorithm [11] builds a surrogate for a likelihood function of latent variable models by using Jensen's inequality. It is also well known for its use in devising an algorithm for NMF [9, 10].

In this section, we derive a majorizer of $\mathcal{J}(U^{(s)},H^{(n)},U^{(n)})$, according to which we obtain the update equations for $U^{(s)}$, $H^{(n)}$ and $U^{(n)}$. First, $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X})$ involves a "log-of-sum" form of $H_{k,\omega}U_{k,t}$. Since the negative logarithm function is a convex function, we can invoke Jensen's inequality to construct an upper bound of $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X})$ with a "sum-of-logs" form in the same way as [9]

$$\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X})\le\mathcal{I}^+(\boldsymbol{Y}|\boldsymbol{X}), \tag{6}$$

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

where $\stackrel{c}{=}$ 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 the equality of Eq. (6) holds if and only if

$$\zeta_{k,\omega,t}=\frac{H_{k,\omega}U_{k,t}}{\sum_{k'}H_{k',\omega}U_{k',t}}. \tag{7}$$

Note that for convenience of notation here we have defined

$$\begin{aligned}H_{k,\omega}&=H^{(s)}_{k,\omega}\ (k=1,\ldots,K_s),\\ H_{k+K_s,\omega}&=H^{(n)}_{k,\omega}\ (k=1,\ldots,K_n),\\ U_{k,t}&=U^{(s)}_{k,t}\ (k=1,\ldots,K_s),\\ U_{k+K_s,t}&=U^{(n)}_{k,t}\ (k=1,\ldots,K_n).\end{aligned}$$

Next, we derive a majorizer of the MFCC-GMM term $-\mathcal{K}(X^{(s)})$. In the same way as Eq. (6), we can construct an upper bound of the negative logarithm function by invoking

Jensen's inequality

$$-\mathcal{K}(X^{(s)}) \leq -\sum_{t,m} \alpha_{m,t} \log \frac{w_m \prod_n \mathcal{N}(\mathcal{X}_{n,t}; \mu_{n,m}, \sigma^2_{n,m})}{\alpha_{m,t}}$$
$$\stackrel{c}{=} \sum_{t,m} \alpha_{m,t} \sum_n \frac{(\mathcal{X}_{n,t} - \mu_{n,m})^2}{2\sigma^2_{n,m}}, \tag{8}$$

where $\alpha_{m,t}$ is also a positive weight that sums to unity. The equality of this inequality holds if and only if

$$\alpha_{m,t} = \frac{w_m \prod_n \mathcal{N}(\mathcal{X}_{n,t}; \mu_{n,m}, \sigma^2_{n,m})}{\sum_{m'} w_{m'} \prod_n \mathcal{N}(\mathcal{X}_{n,t}; \mu_{n,m'}, \sigma^2_{n,m'})}. \tag{9}$$

Since the quadratic function is a convex function, we can employ the inequality used in [12, 13] to construct an upper bound of Eq. (8)

$$(\mathcal{X}_{n,t} - \mu_{n,m})^2 \leq \sum_l \frac{(c_{n,l} \log G_{l,t} - \varphi_{l,n,m,t})^2}{\beta_{l,n,m,t}}, \tag{10}$$

where $\beta_{l,n,m,t}$ is an arbitrary positive number that sums to unity, $\sum_l \beta_{l,n,m,t} = 1$, and $\varphi_{l,n,m,t}$ is a real number that sums to $\mu_{n,m}$, $\sum_l \varphi_{l,n,m,t} = \mu_{n,m}$. For convenience of notation we use $G_{l,t}$ to denote $\sum_\omega f_{l,\omega} X^{(s)}_{\omega,t}$. It can be shown that the equality of this inequality holds if and only if

$$\varphi_{l,n,m,t} = c_{n,l} \log G_{l,t} + \beta_{l,n,m,t}(\mu_{n,m} - \mathcal{X}_{n,t}). \tag{11}$$

From Eqs. (10) and (8), we obtain

$$-\mathcal{K}(X^{(s)}) \leq \sum_{t,l} A_{l,t} (\log G_{l,t})^2 + \sum_{t,l} B_{l,t} \log G_{l,t} + d, \tag{12}$$

where d is a constant term that does not depend on $H_{k,\omega}$ and $U_{k,t}$. Here, for convenience of notation, we have defined

$$A_{l,t} = \sum_{n,m} \frac{\alpha_{m,t} c^2_{n,l}}{2\sigma^2_{n,m} \beta_{l,n,m,t}}, \tag{13}$$

$$B_{l,t} = -\sum_{n,m} \frac{\alpha_{m,t} c_{n,l} \varphi_{l,m,n,t}}{\sigma^2_{n,m} \beta_{l,n,m,t}}. \tag{14}$$

Since $A_{l,t}$ is non-negative, we can use the inequality [14]:

$$(\log G_{l,t})^2 \leq \frac{1}{G_{l,t}} + p(\xi_{l,t}) G_{l,t} + q(\xi_{l,t}), \tag{15}$$

where

$$p(\xi_{l,t}) = \frac{2 \log \xi_{l,t}}{\xi_{l,t}} + \frac{1}{\xi^2_{l,t}}, \tag{16}$$

$$q(\xi_{l,t}) = (\log \xi_{l,t})^2 - 2 \log \xi_{l,t} - \frac{2}{\xi_{l,t}}, \tag{17}$$

to construct an upper bound of the first term of Eq. (12). We can confirm that the equality of this inequality holds if and only if

$$\xi_{l,t} = G_{l,t}. \tag{18}$$

By focusing on the fact that a reciprocal function is convex in the positive domain and that $f_{l,\omega} H_{k,\omega} U_{k,t}$ is positive, we can apply Jensen's inequality to $1/G_{l,t}$

$$\frac{1}{G_{l,t}} = \frac{1}{\sum_{\omega,k} f_{l,\omega} H_{k,\omega} U_{k,t}} \leq \sum_{\omega,k} \frac{\rho^2_{l,k,\omega,t}}{f_{l,\omega} H_{k,\omega} U_{k,t}},$$

where $\rho_{l,k,\omega,t}$ is a positive weight that sums to unity, $\sum_{\omega,k} \rho_{l,k,\omega,t} = 1$. It can be shown that the equality of this inequality holds if and only if

$$\rho_{l,k,\omega,t} = \frac{f_{l,\omega} H_{k,\omega} U_{k,t}}{\sum_{\omega',k'} f_{l,\omega'} H_{k',\omega'} U_{k',t}}. \tag{19}$$

Care must be taken of the fact that $B_{l,t}$ can be either non-negative or negative. Thus, we consider applying different inequalities to the second term of Eq. (12) according to the sign of $B_{l,t}$. First, when $B_{l,t}$ is non-negative, we can show that

$$B_{l,t} \log G_{l,t} \leq B_{l,t} \Big(\frac{G_{l,t}}{\phi_{l,t}} + \log \phi_{l,t} - 1 \Big), \tag{20}$$

where $\phi_{l,t}$ is an arbitrary positive number. This is simply given by the fact that a tangent line to the graph of a differentiable concave function lies entirely above the graph and that a logarithm function is a concave function. We can easily confirm that the equality of this inequality holds if and only if

$$\phi_{l,t} = G_{l,t}. \tag{21}$$

Second, when $B_{l,t}$ is negative, $B_{l,t} \log G_{l,t}$ becomes convex in $G_{l,t}$, and so we can invoke Jensen's inequailty to construct an upper bound

$$B_{l,t} \log G_{l,t} \leq B_{l,t} \sum_{\omega,k} \upsilon_{k,l,\omega,t} \log \frac{f_{l,\omega} H_{k,\omega} U_{k,t}}{\upsilon_{k,l,\omega,t}}, \tag{22}$$

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

$$\upsilon_{k,l,\omega,t} = \frac{f_{l,\omega} H_{k,\omega} U_{k,t}}{\sum_{\omega',k'} f_{l,\omega'} H_{k',\omega'} U_{k',t}}. \tag{23}$$

By combining Eqs. (20) and (22), we can write an upper bound of $B_{l,t} \log G_{l,t}$ as

$$B_{l,t} \log G_{l,t} \leq \delta_{B_{l,t} \geq 0} |B_{l,t}| \Big(\frac{G_{l,t}}{\phi_{l,t}} + \log \phi_{l,t} - 1 \Big)$$
$$- \delta_{B_{l,t} < 0} |B_{l,t}| \sum_{\omega,k} \upsilon_{k,l,\omega,t} \log \frac{f_{l,\omega} H_{k,\omega} U_{k,t}}{\upsilon_{k,l,\omega,t}},$$

where δ is an indicator function that takes the value 1 if its argument is true and 0 otherwise.

To sum up, a majorizer of $-\mathcal{K}(\boldsymbol{X}^{(s)})$ can be written as

$$-\mathcal{K}(X^{(s)}) \tag{24}$$
$$\leq \sum_{t,l} A_{l,t} \Big(\sum_{\omega,k} \frac{\rho^2_{l,k,\omega,t}}{f_{l,\omega} H_{k,\omega} U_{k,t}} + p(\xi_{l,t}) G_{l,t} + q(\xi_{l,t}) \Big)$$
$$+ \sum_{t,l} \delta_{B_{l,t} \geq 0} |B_{l,t}| \Big(\frac{G_{l,t}}{\phi_{l,t}} + \log \phi_{l,t} - 1 \Big)$$
$$- \sum_{t,l} \delta_{B_{l,t} < 0} |B_{l,t}| \sum_{\omega,k} \upsilon_{k,l,\omega,t} \log \frac{f_{l,\omega} H_{k,\omega} U_{k,t}}{\upsilon_{k,l,\omega,t}} + d.$$

By combining this with the majorizer of $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{X})$, we can construct a majorizer of the objective function of interest. This majorizer is particularly noteworthy in that it is given as the sum of a reciprocal function, logarithm functions and a first order function of $H_{k,\omega} U_{k,t}$, which can be minimized analytically with respect to $H_{k,\omega}$ and $U_{k,t}$. We have already seen that the update equations for the auxiliary parameters are given by Eqs. (9), (11), (18), (19), (21) and (23). The next step is to derive the update equations for $H_{k,\omega}$ and $U_{k,t}$.

2.4. Update rules

We can obtain the update rules for $H_{k,\omega}$ and $U_{k,t}$ by setting the partial derivatives of the proposed majorizer with respect to $H^{(s)}$, $U^{(s)}$, $H^{(n)}$ and $U^{(n)}$ at zero. Note that the update rules of the first two are given as the solutions to quadratic equations. Since $H^{(s)}$ and $U^{(s)}$ must be non-negative, a positive solution must be selected. Thus, we finally arrive at the following update equations

$$H^{(s)}_{k,\omega} = \frac{-\mathsf{b}_{k,\omega} + \sqrt{\mathsf{b}^2_{k,\omega} - 4\mathsf{a}_{k,\omega}\mathsf{c}_{k,\omega}}}{2\mathsf{a}_{k,\omega}}, \tag{25}$$

$$H^{(n)}_{k,\omega} = \frac{\sum_t \zeta_{k+K_s,\omega,t} Y_{\omega,t}}{\sum_t U_{k+K_s,t}}, \tag{26}$$

$$U^{(s)}_{k,t} = \frac{-\mathsf{e}_{k,t} + \sqrt{\mathsf{e}^2_{k,t} - 4\mathsf{d}_{k,t}\mathsf{f}_{k,t}}}{2\mathsf{d}_{k,t}}, \tag{27}$$

$$U^{(n)}_{k,t} = \frac{\sum_\omega \zeta_{k+K_s,\omega,t} Y_{\omega,t}}{\sum_\omega H_{k+K_s,\omega}}, \tag{28}$$

where

$$\mathsf{a}_{k,\omega} = \sum_t U_{k,t} + \lambda \sum_{l,t} A_{l,t} p(\xi_{l,t}) f_{l,\omega} U_{k,t} + \lambda \sum_{l,t} \frac{\delta_{B_{l,t}\geq 0}|B_{l,t}|}{\phi_{l,t}} f_{l,\omega} U_{k,t},$$

$$\mathsf{b}_{k,\omega} = -\sum_t \zeta_{k,\omega,t} Y_{\omega,t} - \lambda \sum_{l,t} \delta_{B_{l,t}<0}|B_{l,t}|\upsilon_{k,l,\omega,t}$$

$$\mathsf{c}_{k,\omega} = -\lambda \sum_{l,t} \frac{A_{l,t}\rho^2_{l,k,\omega,t}}{f_{l,\omega}U_{k,t}},$$

$$\mathsf{d}_{k,t} = \sum_\omega H_{k,\omega} + \lambda \sum_{\omega,l} A_{l,t} p(\xi_{l,t}) f_{l,\omega} H_{k,\omega} + \lambda \sum_{\omega,l} \frac{\delta_{B_{l,t}\geq 0}|B_{l,t}|}{\phi_{l,t}} f_{l,\omega} H_{k,\omega},$$

$$\mathsf{e}_{k,t} = -\sum_\omega \zeta_{k,\omega,t} Y_{\omega,t} - \lambda \sum_{\omega,l} \delta_{B_{l,t}<0}|B_{l,t}|\upsilon_{k,l,\omega,t},$$

$$\mathsf{f}_{k,t} = -\lambda \sum_{\omega,l} \frac{A_{l,t}\rho^2_{l,k,\omega,t}}{f_{l,\omega}H_{k,\omega}}.$$

We can confirm that when $\lambda = 0$ the update rules reduce to those of the regular NMF with the I divergence.

3. Experiments

To confirm the effect of the proposed method, we evaluated the cepstral distance (the mean square distance in the MFCC domain) between the estimated and clean speech signals and the signal-to-distortion ratios (SDRs) of the estimated speech signals using the ATR503 database [15]. We chose the conventional semi-supervised NMF method [8] for comparison.

We used four types of noise: white noise, babble noise, measured museum noise and background music noise. The test data were created by adding noise sources to clean speech sources set with different signal-to-noise ratios (SNRs), ranging from -15dB to 5dB. All the audio signals were monaural and sampled at 16kHz. The STFT was computed using a Hanning window that was 32ms long with a 16ms overlap. In the training process, the MFCC vectors were extracted from clean speech samples uttered by 6 female and 4 male speakers (450 sentences for each speaker) and a GMM with 30 components was trained using the EM algorithm. We set the dimension of the MFCC vector at 13. In the separation process, we set λ at 1. The parameters were initialized using a regular semi-supervised NMF that was run for 500 iterations. The enhanced speech signals were obtained using the Wiener filter constructed by the estimated speech and noise spectrograms.

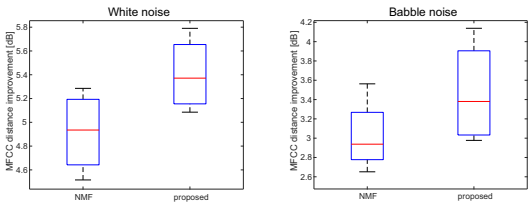

Figure 1: MFCC distance improvement with proposed method and semi-supervised NMF. With white noise (left) and babble noise (right).

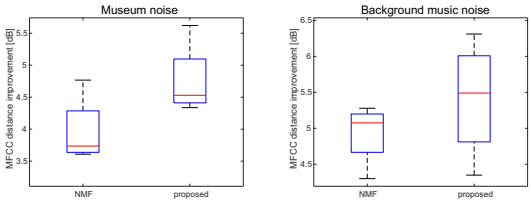

Figure 2: MFCC distance improvement with proposed method and semi-supervised NMF. With measured museum noise (left) and background music noise (right).

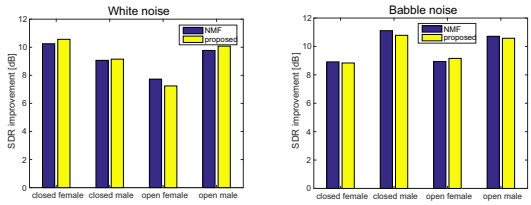

Figure 3: SDR improvement with proposed method and semi-supervised NMF. With white noise (left) and babble noise (right).

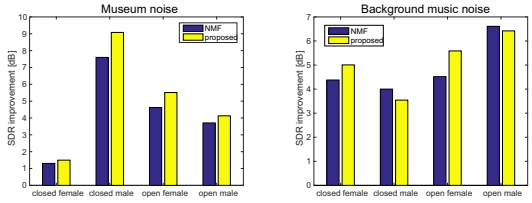

Figure 4: SDR improvement with proposed method and semi-supervised NMF. With measured museum noise (left) and background music noise (right).

Fig. 1–4 show the results of the proposed method tested on four types of noise. As the results show, the proposed method yielded a 0.6 dB higher cepstral distance improvement in average and a slightly higher SDR improvement over conventional semi-supervised NMF.

4. Conclusions

This paper proposed a novel approach for jointly enhancing the spectral and cepstral sequences of noisy speech. The method optimizes the combined objective function consisting of an NMF-based model-fitting criterion defined in the spectral domain and a Gaussian mixture model (GMM)-based probability distribution defined in the cepstral domain based on a majorization-minimization scheme. Experimental results revealed that the proposed method outperformed the conventional NMF approach in terms of both SDR and cepstral distance.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP26730100 and JP26280060, and SECOM Science and Technology Foundation.

6. References

[1] J. Droppo. L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora 2 database," in Proc. Eurospeech 2001, vol. 1, pp. 217–220, 2001.

[2] J. Droppo, L. Deng, and A. Acero, "Evaluation of SPLICE on the Aurora 2 and 3 tasks," in Proc. ICSLP 2002, pp. 29–32, 2002.

[3] K. Chijiiwa, M. Suzuki, N. Minematsu, and K. Hirose, "Unseen noise robust speech recognition using adaptive piecewise linear transformation," in Proc. ICASSP 2012, pp. 4289–4292, 2012.

[4] T. Kristjansson, and J. Hershey, "High resolution signal reconstruction," in Proc. ASRU, pp. 291–296, 2003.

[5] J. Li, M. L. Seltzer, and Y. Gong, "Improvements to VTS feature enhancement," in Proc. ICASSP 2012, pp. 4677–4680, 2012.

[6] S. Watanabe, and J. R. Hershey, "Stereo-based feature enhancement using dictionary learning," in Proc. ICASSP 2013, pp. 7073–7077, 2013.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Trans. ASLP, vol. 23, no. 1, pp. 7–19, 2015.

[8] 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.

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

[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. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of Royal Statistical Society Series B, vol. 39, pp. 1–38, 1977.

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

[13] H. Kameoka, and N. Takamune, "Training restricted Boltzmann machines with auxiliary function approach," in Proc. MLSP 2014, 2014.

[14] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," in Proc. ICASSP 2012, pp. 5365–5368, 2012.

[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.