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

Local Sparsity Based Online Dictionary Learning for Environment-Adaptive Speech Enhancement with Nonnegative Matrix Factorization

Kwang Myung Jeon and Hong Kook Kim

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Korea
{kmjeon, hongkook}@gist.ac.kr

Abstract

In this paper, a nonnegative matrix factorization (NMF)-based speech enhancement method robust to real and diverse noise is proposed by online NMF dictionary learning without relying on prior knowledge of noise. Conventional NMF-based methods have used a fixed noise dictionary, which often results in performance degradation when the NMF noise dictionary cannot cover noise types that occur in real-life recording. Thus, the noise dictionary needs to be learned from noises according to the variation of recording environments. To this end, the proposed method first estimates noise spectra and then performs online noise dictionary learning by a discriminative NMF learning framework. In particular, the noise spectra are estimated from minimum mean squared error filtering, which is based on the local sparsity defined by a posteriori signal-to-noise ratio (SNR) estimated from the NMF separation of the previous analysis frame. The effectiveness of the proposed speech enhancement method is demonstrated by adding six different realistic noises to clean speech signals with various SNRs. Consequently, it is shown that the proposed method outperforms comparative methods in terms of signal-to-distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) for all kinds of simulated noise and SNR conditions.

Index Terms: speech enhancement, diverse noise, environment adaptation, nonnegative matrix factorization, online dictionary learning, local sparsity

1. Introduction

Recently, speech enhancement has become more demanding because speech-based applications such as speech communication and automatic speech recognition are mostly operated in diverse noisy environments [1, 2]. In order to enhance speech signals recorded under such noise conditions, noise spectral components should be suppressed without damaging spectral components belonging to the target speech signal [3]. To this end, conventional efforts have been focused on estimating noise power spectra from the noisy signal [4, 5] or separating speech and noise from the noisy speech [6–15]

Among the successful noise estimation methods is the minima-controlled recursive algorithm (MCRA) [4, 5]. MCRA estimates noise by only tracking minimum statistics in noise-only regions. That is, noise-only regions are detected when the ratio between the noisy speech power spectrum and the minimum power spectrum is below a pre-defined threshold. However, the main drawback of MCRA is that non-stationary noises, such as harmonic or tonal noises, are difficult to estimate because their sparse characteristics in time and/or frequency are not suitable to be modelled by using only minimum statistics [6]. Thus, speech enhancement methods for use with challenging real environmental noises should consider both stationary and non-stationary characteristics of noise.

As an alternative to the minima tracking based noise estimation approaches, NMF-based source separation has been successfully realized into speech enhancement [7–15]. Several studies reported that NMF was suitable for separating speech signals from interfering non-stationary noise such as wind or television noise [7, 8]. In general, the separation performance of the NMF-based methods is guaranteed if speech and noise NMF dictionaries are trained sufficiently to represent arbitrary noisy spectral magnitudes [7–9]. Thus, many NMF-based speech enhancement methods have trained speech and noise dictionaries in advance by using extensive speech and noise databases, respectively. Recently, discriminative NMF training was introduced to reduce ambiguities between speech and noise dictionaries by retraining their bases for a given noisy spectra and activations, resulting improved speech separation performance under known noise conditions [15]. However, the noise dictionary cannot be always prepared in advance, because the existing noise database for training the noise dictionary represents only specific types of noise among infinite possible noise types [10]. This limitation leads to the performance degradation of the NMF-based methods when the types of noise between training and evaluation steps of NMF are mismatched [11].

To alleviate this problem, several previous works sought to semi-supervised NMF-based source separation techniques, which directly learn noise dictionary at the separation step of NMF [11, 12]. In other words, the semi-supervised approaches measure a mismatch between the fixed speech dictionary and the observed noisy spectral magnitudes, and then they incorporate the mismatch into the update rule of the noise dictionary. However, such semi-supervised NMF approaches are apt to fail in separating speech from noise when their spectral distributions are excessively overlapped, which generally happens when speech signals are recorded in real-world noisy environments. This is because most of the overlapped spectral components are updated as if they were noise spectral components in the semi-supervised NMF process [12].

To mitigate this drawback of the semi-supervised NMF approaches, this paper proposes an NMF-based environment-adaptive speech enhancement method. Unlike the semi-supervised NMF techniques, the proposed method estimates noise spectral magnitudes at the enhancement stage of the previous analysis frame, and then utilizes them for noise dictionary learning on the fly. Specifically, the proposed method decomposes noisy spectral magnitudes into speech and noise magnitudes by using a supervised NMF technique. Then, degree of overlap between the separated speech and noise, which is called *local*

http://dx.doi.org/10.21437/Interspeech.2016-586

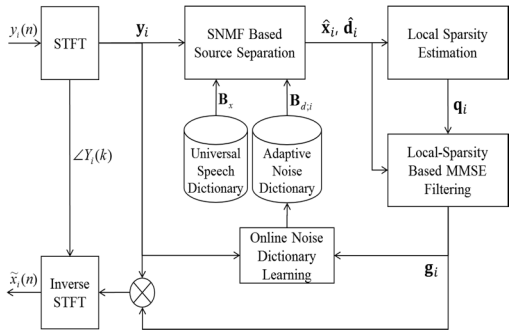

Figure 1: *Procedure of the proposed speech enhancement method.*

sparsity, is estimated by measuring the *a posteriori* signal-to-noise ratio (SNR) for each frequency bin. Next, the estimated local sparsity is incorporated into the construction of a minimum mean squared error (MMSE) filter in order to obtain the noise component for noise dictionary learning, and a discriminative NMF learning procedure [15] is applied for the noise dictionary learning. Finally, the learnt noise dictionary is recursively fed into the NMF separation for the following noisy frame.

2. Proposed Local-Sparsity Based Online Dictionary Learning

Figure 1 shows the procedure of the proposed speech enhancement method based on local sparsity estimation and online noise dictionary learning. Like the conventional NMF-based speech enhancement methods [8, 9, 13], the proposed method first decomposes spectral magnitude of noisy speech signal at the *i*-th frame, $\mathbf{y}_i$, into those of speech and noise, $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{d}}_i$, by using the supervised sparse NMF technique [13] with a fixed speech dictionary, $\mathbf{B}_x$, and an adaptive noise dictionary, $\mathbf{B}_{d;i}$. Subsequently, the local sparsity is calculated at every frequency bin by using a ratio between $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{d}}_i$, and then it is plugged into constructing an MMSE filter for both speech enhancement and online noise dictionary learning for the (*i*+1)-th frame. The following subsections describe each step of the proposed method in detail.

2.1. NMF-Based Speech and Noise Separation

First of all, the *n*-th speech sample at the *i*-th speech frame, $y_i(n)$, is represented as

$$y_i(n) = x_i(n) + d_i(n) \tag{1}$$

where $x_i(n)$ and $d_i(n)$ are clean speech and additive noise at the *i*-th frame, respectively. Note that $d_i(n)$ is assumed to be uncorrelated with $x_i(n)$. By applying a *K*-point short-time Fourier transform (STFT) to (1), $y_i(n)$ can be represented in the frequency domain as

$$Y_i(k) = X_i(k) + D_i(k) \quad \text{for} \quad k = 0,1,\cdots,K-1 \tag{2}$$

where $Y_i(k)$, $X_i(k)$, and $D_i(k)$ denote the *k*-th spectral components of $y_i(n)$, $x_i(n)$, and $d_i(n)$, respectively. To separate $X_i(k)$ and $D_i(k)$ from $Y_i(k)$, the *p*-powered spectral magnitude of noisy speech frame is represented as $|Y_i(k)|^p \cong |X_i(k)|^p + |D_i(k)|^p$, according to satisfactory results of NMF-based noise reduction when p is 1 or 2 [7–14]. For simplicity, $|Y_i(k)|^p$, $|X_i(k)|^p$, and $|D_i(k)|^p$ are represented as $\mathbf{y}_i$, $\mathbf{x}_i$, and $\mathbf{d}_i$, which are all $K \times 1$ matrices.

In the NMF framework, $\mathbf{y}_i = \mathbf{B}_y\mathbf{a}_{y;i}$, $\mathbf{x}_i = \mathbf{B}_x\mathbf{a}_{x;i}$, and $\mathbf{d}_i = \mathbf{B}_{d;i}\mathbf{a}_{d;i}$, respectively, where $\mathbf{B}_y$, $\mathbf{B}_x$, and $\mathbf{B}_{d;i}$ are the dictionaries of $\mathbf{y}_i$, $\mathbf{x}_i$, and $\mathbf{d}_i$, respectively. Moreover, $\mathbf{a}_{y;i}$, $\mathbf{a}_{x;i}$, and $\mathbf{a}_{d;i}$ are the activation matrices corresponding to $\mathbf{B}_y$, $\mathbf{B}_x$, and $\mathbf{B}_{d;i}$ at the *i*-th frame, respectively. By assuming that $\mathbf{x}_i$ and $\mathbf{d}_i$ are fully separable from $\mathbf{y}_i$, $\mathbf{y}_i$ can be rewritten as [8]

$$\mathbf{y}_i = \mathbf{B}_y\mathbf{a}_{y;i} = \begin{bmatrix}\mathbf{B}_x\mathbf{B}_{d;i}\end{bmatrix}\begin{bmatrix}\mathbf{a}_{x;i}\\ \mathbf{a}_{d;i}\end{bmatrix} = \mathbf{B}_x\mathbf{a}_{x;i} + \mathbf{B}_{d;i}\mathbf{a}_{d;i} \tag{3}$$

where $\mathbf{B}_y = [\mathbf{B}_x\mathbf{B}_d]$ and $\mathbf{a}_{y;i} = [\mathbf{a}_{x;i}\mathbf{a}_{d;i}]^T$. Note that T refers to the transpose operation. If R_x and R_d ($R_y = R_x + R_d$) are the ranks of the dictionaries for $\mathbf{x}_i$ and $\mathbf{d}_i$, respectively, then the dimensions of $\mathbf{B}_y$, $\mathbf{B}_x$, and $\mathbf{B}_{d;i}$ are $K \times R_y$, $K \times R_x$, and $K \times R_d$, respectively, while the dimensions of $\mathbf{a}_{y;i}$, $\mathbf{a}_{x;i}$, and $\mathbf{a}_{d;i}$ are $R_y \times 1$, $R_x \times 1$, and $R_d \times 1$, respectively.

Since supervised NMF-based speech enhancement methods assume that both $\mathbf{B}_x$ and $\mathbf{B}_{d;i}$ are given in advance [7, 8, 13], they focus on finding $\mathbf{a}_{x;i}$ and $\mathbf{a}_{d;i}$ from $\mathbf{y}_i$ for the separation of speech and noise. To achieve this goal, a multiplicative update rule with a sparsity constraint [13] is iteratively performed as

$$\begin{bmatrix}\mathbf{a}_{x;i}^j\\ \mathbf{a}_{d;i}^j\end{bmatrix} = \begin{bmatrix}\mathbf{a}_{x;i}^{j-1}\\ \mathbf{a}_{d;i}^{j-1}\end{bmatrix} \otimes \frac{\begin{bmatrix}\mathbf{B}_x\mathbf{B}_{d;i}\end{bmatrix}^T \dfrac{\mathbf{y}_i}{\begin{bmatrix}\mathbf{B}_x\mathbf{B}_{d;i}\end{bmatrix}\begin{bmatrix}\mathbf{a}_{x;i}^{j-1}\mathbf{a}_{d;i}^{j-1}\end{bmatrix}^T}}{\begin{bmatrix}\mathbf{B}_x\mathbf{B}_{d;i}\end{bmatrix}^T\mathbf{1} + \boldsymbol{\mu}} \tag{4}$$

where j is an iteration index and $\boldsymbol{\mu}$ is an $R_y \times 1$ matrix in which all elements are equal to a sparsity weight of the ℓ_1 constraint, which is set to 5 according to the previous work [13]. In addition, $\otimes$ and / indicate element-wise multiplication and division, respectively. Moreover, $\mathbf{1}$ in (4) is a $K \times 1$ matrix in which all elements are equal to unity. Note that all elements of $\mathbf{a}_{y;i}^0 = \begin{bmatrix}\mathbf{a}_{x;i}^0\mathbf{a}_{d;i}^0\end{bmatrix}^T$ can be initialized as random values between 0 and 1 [13]. In NMF separation, (4) is repeated until the relative reduction of an NMF objective function is less than a pre-defined threshold. In this paper, the Kullback–Leibler (KL) divergence is employed as an NMF objective function [8, 13, 16]. Consequently, the separated spectral magnitude of speech and noise at the *i*-th frame, $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{d}}_i$, respectively, are obtained as $\hat{\mathbf{x}}_i = \mathbf{B}_x\mathbf{a}_{x;i}^J$ and $\hat{\mathbf{d}}_i = \mathbf{B}_{d;i}\mathbf{a}_{d;i}^J$, where J is the iteration to make (4) converged.

2.2. Local Sparsity Estimation

In order to improve separability between speech and noise, a local sparsity is estimated. To this end, a local SNR at the *k*-th frequency band, $r_i(k)$, is defined as

$$r_i(k) = \frac{\hat{x}_i(k)}{\hat{d}_i(k)} \tag{5}$$

where $\hat{x}_i(k)$ and $\hat{d}_i(k)$ are the spectral magnitude of the separated speech and noise at the *k*-th frequency band, respectively. Next, the local sparsity at the *i*-th frame and *k*-th frequency bin is measured by the ℓ_1/ℓ_2 sparsity-inducing norm [16] of the local SNR over all the K_q adjacent frequency bins of the past I_q frames, such as

$$q_i(k) = \frac{1}{\sqrt{N}-1}\left(\sqrt{N} - \frac{\sum_{j=i-I_q}^{i}\sum_{l=k-\left\lfloor\frac{1}{2}K_q\right\rfloor}^{k+\left\lfloor\frac{1}{2}K_q\right\rfloor}\bar{r}_j(l)}{\sqrt{\sum_{j=i-I_q}^{i}\sum_{l=k-\left\lfloor\frac{1}{2}K_q\right\rfloor}^{k+\left\lfloor\frac{1}{2}K_q\right\rfloor}\bar{r}_j^2(l)}}\right) \tag{6}$$

where $\lfloor x \rfloor$ is the integer smaller than or equal to x, $N = (K_q + 1)(I_q + 1)$, and $\bar{r}_j(l) = r_j(l)/\max_{k\in[0,K-1]}\{r_j(k)\}$. Note that $q_i(k)$ becomes one if the distribution of the local SNRs around the i-th frame and k-th frequency bin are highly sparse, whereas densely distributed local SNRs lead $q_i(k)$ to zero. In other words, $q_i(k)$ becomes larger as $\hat{x}_i(k)$ and $\hat{d}_i(k)$ are sparsely separated, which demonstrates that $q_i(k)$ can be considered as a confidence metric for the separation at the k-th frequency bin. Therefore, $q_i(k)$ is plugged for constructing an MMSE filter to improve separability, which will be explained in the next subsection.

2.3. Local-Sparsity Based MMSE Filtering

In this subsection, a noise reduction filter based on the MMSE criteria with local sparsity is constructed for both the speech enhancement and the noise estimation for online noise dictionary learning. To this end, the *a priori* SNR, $\boldsymbol{\xi}_i$, is first estimated using $\hat{\mathbf{x}}_i$, $\hat{\mathbf{d}}_i$, and the local-sparsity, $\mathbf{q}_i$, with a decision-directed approach. That is,

$$\boldsymbol{\xi}_i = \frac{\alpha\tilde{\mathbf{x}}_{i-1} + (1-\alpha)\hat{\mathbf{x}}_i \otimes \mathbf{q}_i}{\bar{\mathbf{d}}_{i-1}} \tag{7}$$

where $\mathbf{q}_i$ is a $K \times 1$ matrix consisting of $q_i(k)$; α is a smoothing coefficient for the decision-directed $\boldsymbol{\xi}_i$, and it is set to 0.3 empirically. In addition, $\bar{\mathbf{d}}_i$ in (7) is a time-smoothed version of $\hat{\mathbf{d}}_i$, and it is realized as

$$\bar{\mathbf{d}}_i = \gamma\bar{\mathbf{d}}_{i-1} + \beta_i(1-\gamma)\hat{\mathbf{d}}_i \tag{8}$$

where $\bar{\mathbf{d}}_0 = \hat{\mathbf{d}}_1$, and γ controls the stationarity of $\bar{\mathbf{d}}_i$ and is set to 0.85. In (8), β_i is an adaptive noise flooring factor at the i-th frame, which is derived from the ratio between the normalized activation powers of separated noise and speech, as

$$\beta_i = 20\log_{10}\frac{R_x\sum_{r=1}^{R_d} a_{d;i}^J(r)}{R_d\sum_{r=1}^{R_x} a_{x;i}^J(r)} \tag{9}$$

where $a_{x;i}^J(r)$ and $a_{d;i}^J(r)$ indicate an r-th element of $\mathbf{a}_{x;i}^J$ and $\mathbf{a}_{d;i}^J$ from (4), respectively. Next, an MMSE filter is constructed as

$$\mathbf{g}_i = \frac{\boldsymbol{\xi}_i}{\mathbf{1} + \boldsymbol{\xi}_i} \tag{10}$$

and an enhanced speech spectral magnitude, $\tilde{\mathbf{x}}_i$, is obtained by applying (10) to $\mathbf{y}_i$; thus $\tilde{\mathbf{x}}_i = \mathbf{g}_i \otimes \mathbf{y}_i$.

Finally, an enhanced speech signal at the i-th frame, $\tilde{x}_i(n)$, is obtained by applying an inverse STFT to $|\tilde{X}_i(k)|$, which is an element of $\tilde{\mathbf{x}}_i$, with the phase of the input signal, $\angle Y_i(k)$.

2.4. Discriminative Noise Dictionary Learning

For the noise dictionary learning in a discriminative way, a reference noise spectral magnitude for the noise dictionary learning, $\tilde{\mathbf{d}}_i$, is estimated by using the local-sparsity based MMSE filter, $\mathbf{g}_i$, as described in (10). That is, $\tilde{\mathbf{d}}_i$ is estimated only when the noise activation is dominant, such as

$$\tilde{\mathbf{d}}_i = \begin{cases} \mathbf{y}_i \otimes (1-\mathbf{g}_i), & if \quad \beta_i > Q_i \\ \bar{\mathbf{d}}_i, & otherwise \end{cases} \tag{11}$$

where Q_i controls whether or not the reference noise should be updated according to the local sparsity, which is defined as

$$Q_i = 20\log_{10}\frac{K}{\epsilon(K - \sum_{k=0}^{K-1} q_i(k))} \tag{12}$$

where $\epsilon$is an adjustment factor to control the effect of sparsity on noise update in (11) and it is set to 0.8 by exhaustive experiments. Next, M frames of $\tilde{\mathbf{d}}_i$ and $\mathbf{a}_{d;i}^J$ are stacked as $\widetilde{\mathbf{D}}_i = [\tilde{\mathbf{d}}_{i-M+1}\cdots\tilde{\mathbf{d}}_i]$ and $\mathbf{A}_{d;i} = [\mathbf{a}_{d;i-M+1}^J\cdots\mathbf{a}_{d;i}^J]$, respectively, where M is set to 10.

In this work, each noise basis is tested for whether it should be updated by

$$I(r) = \begin{cases} 1, & if \quad \left(\frac{Q_i}{M}\sum_{\hat{\imath}=i-M+1}^{i} a_{d;\hat{\imath}}(r)\right) > \bar{A} \\ 0, & otherwise \end{cases} \tag{13}$$

where $\bar{\mathrm{A}} = (\sum_{r=1}^{R_x} a_{x;i}^J(r))/R_x$ and $I(r) = 1$ means that the r-th basis should be updated to accommodate the noise appeared at the i-th frame. Then, $\mathbf{A}_{d;i}$ is decomposed depending on (13) into $\mathbf{A}_{d;i}^{r\in\boldsymbol{I}_u}$ and $\mathbf{A}_{d;i}^{r\in\boldsymbol{I}_f}$, where $\boldsymbol{I}_u = \{r|I(r)=1\}$ and $\boldsymbol{I}_f = \{r|I(r)=0\}$. By using $\widetilde{\mathbf{D}}_i$ and $\mathbf{A}_{d;i}^{r\in\boldsymbol{I}_u}$, the learnt noise dictionary for the (i+1)-th frame, $\widehat{\mathbf{B}}_{d;i+1}^j$, is iteratively updated by minimizing the KL divergence by applying the discriminative dictionary learning technique [15] as

$$\widehat{\mathbf{B}}_{d;i+1}^j = \widehat{\mathbf{B}}_{d;i+1}^{j-1} \otimes \frac{\widetilde{\mathbf{D}}_i}{\dfrac{\widehat{\mathbf{B}}_{d;i+1}^{j-1}\left(\mathbf{A}_{d;i}^{r\in\boldsymbol{I}_u}\right)^T}{\mathbf{1}\left(\mathbf{A}_{d;i}^{r\in\boldsymbol{I}_u}\right)^T}} \tag{14}$$

where $\widehat{\mathbf{B}}_{d;i+1}^0 = \mathbf{B}_{d;i}^{r\in\boldsymbol{I}_u}$ and j is an iteration index. Finally, $\mathbf{B}_{d;i+1}$ is obtained by concatenating the converged $\widehat{\mathbf{B}}_{d;i+1}^{j^*}$ and fixed noise dictionary, $\mathbf{B}_{d;i}^{r\in\boldsymbol{I}_f}$, as $\mathbf{B}_{d;i+1} = [\widehat{\mathbf{B}}_{d;i+1}^{j^*}\ \mathbf{B}_{d;i}^{r\in\boldsymbol{I}_f}]$, which will be used for the NMF-based speech and noise separation for the next frame.

3. Performance Evaluation

The performance of the proposed speech enhancement method was evaluated by measuring the signal-to-distortion ratio (SDR), signal-to-interference-ratio (SIR), signal-to-artifact ratio (SAR) [17], and perceptual evaluation of speech quality (PESQ) [18]. In addition, it was compared with those of four different conventional speech enhancement methods, including the improved minima-controlled recursive averaging (IMCRA) [5], sparse NMF [13], NMF with exemplar dictionary [14], and semi-supervised NMF [11].

In order to training speech and noise dictionaries for NMF-based approaches, clean speech and background noise signals of 12 minutes long each were excerpted from the training set of the CHiME3 corpus [19] and NOISEX-92 database [20], respectively. Note that the ranks of the speech and noise dictionaries for the proposed method as well as sparse and semi-supervised NMF, R_x and R_d, respectively, were all 100, while they were set 1,000 for NMF with exemplar dictionary. In particular, K and p were commonly set to 513 and 2, respectively, for all the NMF-based methods. Moreover, K_q and I_q for the local sparsity estimation of the proposed method were set to 60 and 10, respectively.

For the objective evaluation, 20 speech clips from the TIMIT corpus [21] were prepared, including 13 male and 7 female speech clips, the average length of which was 3.6 seconds long. These speech clips were then artificially added with each of six different environmental noises from the DEMAND database [22] under different SNR conditions ranging from 0 to 15 dB at a step of 5 dB (categories of noises were DLIVING, NRIVIER,

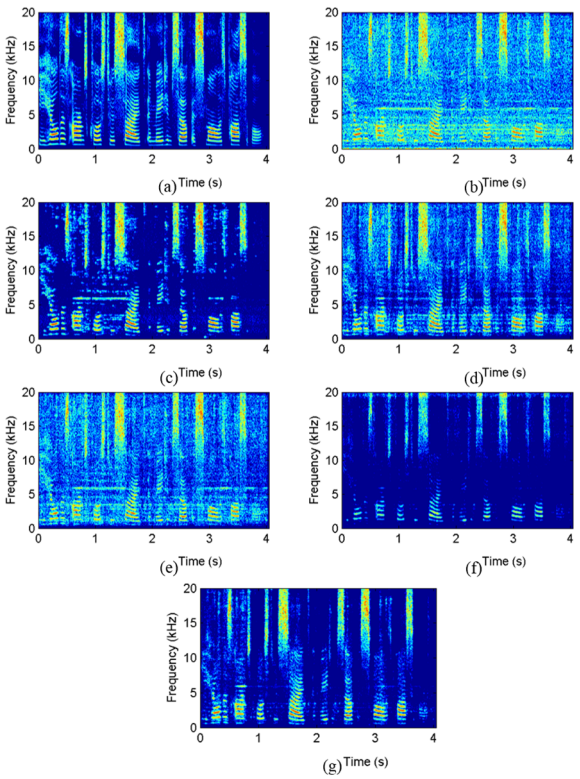

Figure 2: *Spectrograms of (a) a clean speech signal, (b) a noisy speech signal added with train noise at 0 dB SNR; enhanced speech signals by (c) IMCRA, (d) sparse NMF, (e) supervised NMF with exemplar dictionaries, (f) semi-supervised NMF, and (g) the proposed method.*

OOFFICE, PCAFETER, STRAFFIC, and TMETRO). Consequently, 120 noisy speech clips were generated in total, and the proposed method and four conventional ones were applied to enhance such noisy speech clips. In this study, all the speech and noise signals were sampled at 16 kHz with 16-bit resolution.

Figure 2 compares the spectrograms of speech signals enhanced by the proposed and conventional methods. Here, a sample speech clip was mixed with a train noise (TMETRO) at 0 dB SNR. By comparing Figs. 2(a) and 2(c), IMCRA generated excessive spectral peak components at around 1 s, which could be perceived as musical noises. On the other hand, such musical noise was mitigated by applying conventional NMFs, as shown in Figs. 2(d)–2(f), while sparse NMF and supervised NMF with exemplars failed to reduce background noise. In addition, the semi-supervised NMF notably reduced noises; thus, it distorted speech components. Compared to conventional methods, the proposed method suppressed background noises while enhancing the speech signal with much less distortion than other methods. This implies that the proposed method could provide better speech quality in severe noisy environments than conventional methods. Note here that for the proposed method, average number of updated noise bases in (13) was measured as 40 and the dictionary learning of (14) was finished within 20 iterations on average.

Next, Figure 3 compares average SDR, SIR, SAR, and PESQ for all the test noisy speech clips for the proposed method with those of the four conventional methods. In the figure, the vertical line at the top of each bar denotes the standard deviation for a statistical analysis. As shown in the figure, among all the methods, the proposed method had the highest SDR, SIR, SAR,

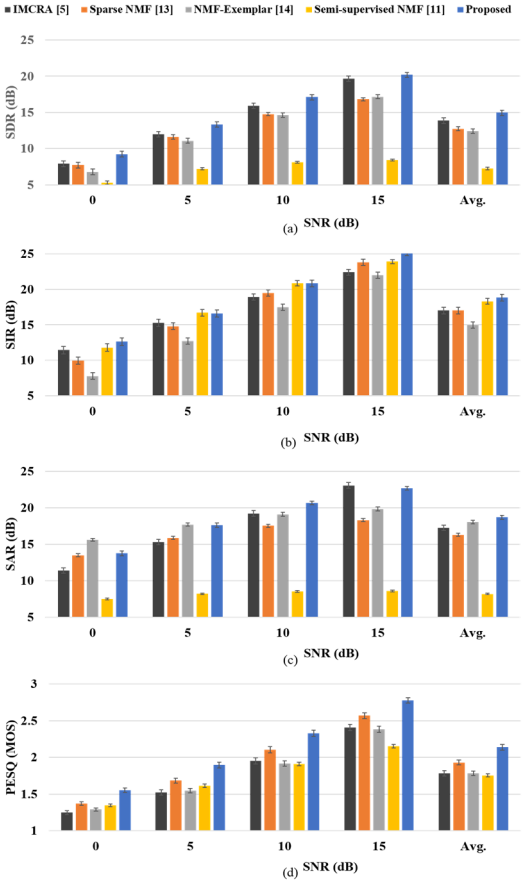

Figure 3: *Comparison of objective quality measures for different speech enhancement methods under various SNR conditions; (a) SDR, (b) SIR, (c) SAR, and (d) PESQ.*

and PESQ for all SNRs. In particular, the proposed method significantly improved average PESQ score by 0.30 and 0.21 dB, compared with IMCRA and semi-supervised NMF, respectively, even under the severe noise condition at 0 dB SNR.

4. Conclusion

In this paper, a local-sparsity based noise dictionary update was proposed for the NMF-based speech enhancement under realistic noise conditions. The proposed method estimated local sparsity from the *a posteriori* SNR to improve the separability at the enhancement step. In addition, the estimated local sparsity was incorporated into online noise dictionary learning, which made the proposed method robust to diverse noises. The performance of the proposed method was compared with those of several conventional methods such as IMCRA, sparse NMF, supervised NMF with exemplars, and semi-supervised NMF. It was shown from the comparison that the proposed method outperformed the conventional methods in terms of SDR, SIR, and PESQ without hurting SAR at different noise types and SNRs.

5. Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (No. 2015R1A2A1A05001687), and by the MSIP, Korea, under the ITRC support program (IITP-2016-H8501-16-1016) supervised by the IITP.

6. References

[1] T. Gerkmann, M. K. Becker, and J. Le-Roux, "Phase processing for single-channel speech enhancement: history and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[2] J. Li, L. Deng, Y. Gong, and R. H. Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[3] P. C. Loizou, *Speech Enhancement – Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[4] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[6] K. M. Jeon, N. I. Park, H. K. Kim, M. K. Choi, and K. I. Hwang, "Mechanical noise suppression based on non-negative matrix factorization and multi-band spectral subtraction for digital cameras," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 2, pp. 296–302, 2013.

[7] M. N. Schmidt, J. Larsen, and F. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Proceedings of IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Thessaloniki, Greece, 2007, pp. 431-436.

[8] K. M. Jeon, H. K. Kim, S. J. Lee, and Y. K. Lee, "Nonnegative matrix factorization based adaptive noise sensing over wireless sensor networks," *International Journal of Distributed Sensor Networks*, doi:10.1155/2014/ 640915, 2014.

[9] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, 2013, pp. 141-145.

[10] N. Mohammadiha, P. Smargadis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[11] F. Weninger, M. Woellmer, and B. Schuller, "Sparse, hierarchical and semi-supervised base learning for monaural enhancement of conversational speech," in *Proceedings of ITG Symposium on Speech Communication*, Braunschweig, Germany, 2012, pp. 1-4.

[12] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA ICA),* Tel Aviv, Israel, 2012, pp. 322-329.

[13] J. Le Roux, F. J. Weninger, and J. R. Hershey, *Sparse NMF–half-baked or Well Done?,* Mitsubishi Electric Research Labs (MERL), Cambridge, MA, Tech. Rep. TR-2015-23, 2015.

[14] J. F. Gemmeke, T. Virtanen, A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[15] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Proceedings of Interspeech*, Singapore, 2014, pp. 865–869.

[16] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research,* vol. 5, no. 1, pp. 1457–1469, 2004.

[17] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 4, pp. 1462–1469, 2006.

[18] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ), and Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, 2000.

[19] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ, pp. 1-8, 2015.

[20] A. Varga and H. J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[21] J. S. Garofolo , L. F. Lamel , W. M. Fisher , J. G. Fiscus , D. S. Pallett and N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, 1993.

[22] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings," *Journal of the Acoustic Society of America*, vol. 133, no. 5, p. 3591, 2013.