Non-negative Matrix Factorization Based Noise Reduction for Noise Robust Automatic Speech Recognition

Seon Man Kim[1], Ji Hun Park[1], Hong Kook Kim[1,*],
Sung Joo Lee[2], and Yun Keun Lee[2]

[1] School of Information and Communications
Gwangju Institute of Science and Technology, Gwangju 500-712, Korea
{kobem30002,jh_park,hongkook}@gist.ac.kr
[2] Speech/Language Information Research Center
Electronics and Telecommunications Research Institute, Daejeon 305-700, Korea
{lee1862,yklee}@etri.re.kr

Abstract. In this paper, we propose a noise reduction method based on non-negative matrix factorization (NMF) for noise-robust automatic speech recognition (ASR). Most noise reduction methods applied to ASR front-ends have been developed for suppressing background noise that is assumed to be stationary rather than non-stationary. Instead, the proposed method attenuates non-target noise by a hybrid approach that combines a Wiener filtering and an NMF technique. This is motivated by the fact that Wiener filtering and NMF are suitable for reduction of stationary and non-stationary noise, respectively. It is shown from ASR experiments that an ASR system employing the proposed approach improves the average word error rate by 11.9%, 22.4%, and 5.2%, compared to systems employing the two-stage mel-warped Wiener filter, the minimum mean square error log-spectral amplitude estimator, and NMF with a Wiener post-filter , respectively.

Keywords: Automatic speech recognition (ASR), Non-negative matrix factorization (NMF), Noise reduction, Non-stationary background noise, Wiener filter.

1 Introduction

Most automatic speech recognition (ASR) systems often suffer considerably from unexpected background noise [1]. Thus, many noise-robust methods in the frequency domain have been reported such as spectral subtraction [2], minimum mean square error log-spectral amplitude (MMSE-LSA) estimation [3], and Wiener filtering [4][5]. In general, conventional front-ends employing such noise reduction methods perform well in stationary noise environments but not always in non-stationary ones. This is because noise reduction is usually performed by estimating noise components during

* This work was supported in part by the ISTD Program, 10035252, and by the Basic Science Research Program through the NRF of Korea funded by the MOST, 2011-0026201.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 338–346, 2012.

the period when target speech is declared inactive under the stationary noise assumption [1][4].

On the other hand, a non-negative matrix factorization (NMF) technique [6] can provide an alternative to estimate target speech from an observed noisy signal. However, the performance of noise reduction methods based on NMF might be degraded when speech and noise have similar distributions in the frequency domain [7]. In other words, there is a large overlap between speech and noise in the frequency domain, thus a certain degree of residual noise remains in the estimated target speech while some speech components are apt to be missed in the target speech. To overcome this problem, we have proposed an NMF-based target speech enhancement method [8], where a Wiener filter was applied to a weighted-sum of speech bases in order to remove the residual noise from the estimated speech. In particular, the temporal continuity constraint technique [9] was also employed so that the characteristics of residual noise remained in the estimated NMF-based target speech became stationary. On the other hand, the target speech was a little damaged after the NMF procedure, even though a regularization technique [7] had been used. Therefore, we need to mitigate such a problem.

In order to mitigate the problem mentioned above, we propose a noise reduction method based on non-negative matrix factorization (NMF) and apply it to noise-robust ASR. The proposed method attenuates non-target noise by a hybrid approach that combines a Wiener filtering and an NMF technique. In addition, stationary noise is estimated from recursively averaging noise components during inactive speech intervals. On the other hand, non-stationary noise is estimated as the difference between the original noise and the estimated noise variance based on recursive averaging. After that, the estimated stationary and non-stationary noises are reduced by Wiener filtering and NMF, respectively. Note here that the NMF bases of the non-stationary noise are trained using a non-stationary noise database (DB), which is generated from an original noise DB.

The rest of this paper is organized as follows. Section 2 proposes an NMF-based noise reduction method. Section 3 demonstrates the effect of the proposed method on ASR performance, and Section 4 concludes this paper.

2 Proposed NMF-Based Noise Reduction Method for ASR

Fig. 1 shows an overall procedure of the proposed noise reduction method which combines NMF with a conventional Wiener filter. As shown in the figure, in the training stage the speech and non-stationary noise bases, $\overline{\mathbf{B}}_S$ and $\overline{\mathbf{B}}_D$, are estimated from speech and non-stationary noise databases (DBs), $\overline{\mathbf{S}}$ and $\overline{\mathbf{D}}$, respectively. In particular, the non-stationary noise DB, $\overline{\mathbf{D}}$, is obtained by applying a recursive averaging method [10] to the original noise DB, $\overline{\mathbf{Y}}$. Note that $\overline{\mathbf{S}}$ or $\overline{\mathbf{Y}}$ is represented in a matrix form by concatenating a sequence of absolute values of speech or noise spectra along the analysis frame, respectively. In the noise reduction stage, activation matrices of target speech and non-stationary (or residual) noise, $\mathbf{A}_S$ and $\mathbf{A}_D$, are estimated by an NMF multiplicative updating rule in order to approximate Wiener

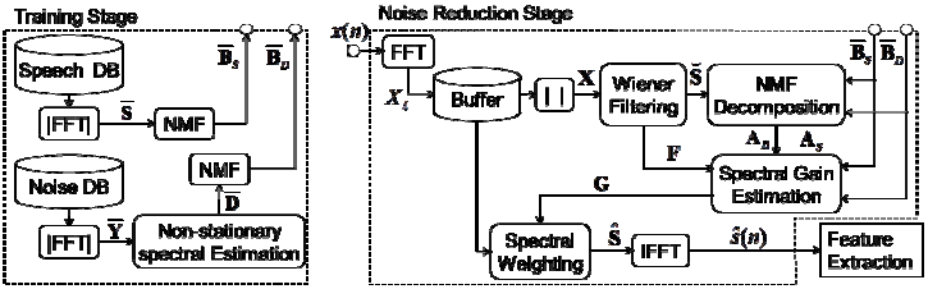

Fig. 1. Block diagram of the proposed NMF-based noise reduction technique applied as a front-end of ASR

filtered outputs, $\breve{\mathbf{S}}$, using known $\overline{\mathbf{B}}_S$ and $\overline{\mathbf{B}}_D$. Then, the weighting value matrix for noise spectral attenuation, $\mathbf{G}$, is obtained from the Wiener filter coefficient matrix, $\mathbf{F}$, and the NMF decomposition outputs, $\mathbf{A}_S$, $\mathbf{A}_D$, $\overline{\mathbf{B}}_S$, and $\overline{\mathbf{B}}_D$. Next, target speech spectral components, $\hat{\mathbf{S}}$, are obtained from $\mathbf{G}$. After that, target speech spectral components are transformed into a time-domain signal, $\hat{s}(n)$, by using an overlap-add method, and $\hat{s}(n)$ is finally used for mel-frequency cepstral coefficient (MFCC) extraction for ASR.

Let $x(n)$, $s(n)$, and $y(n)$ be noisy speech, target speech, and additive noise, respectively, where $x(n)$ and $y(n)$ are assumed to be uncorrelated. In addition, we have $X_k(\ell) = S_k(\ell) + Y_k(\ell)$, where $X_k(\ell)$, $S_k(\ell)$, and $Y_k(\ell)$ denote the spectral components of $x(n)$, $s(n)$, and $y(n)$, respectively, at the k-th frequency bin index $(k = 0,1,\cdots,K-1)$ and ℓ-th segmented frame index $(\ell = 0,1,2,\cdots)$.

As mentioned in Section 1, the performance of an NMF-based noise reduction method could be degraded when speech and noise have similar distributions in the frequency domain [8]. Figs. 2(a) and 2(b) show the spectral distribution and basis

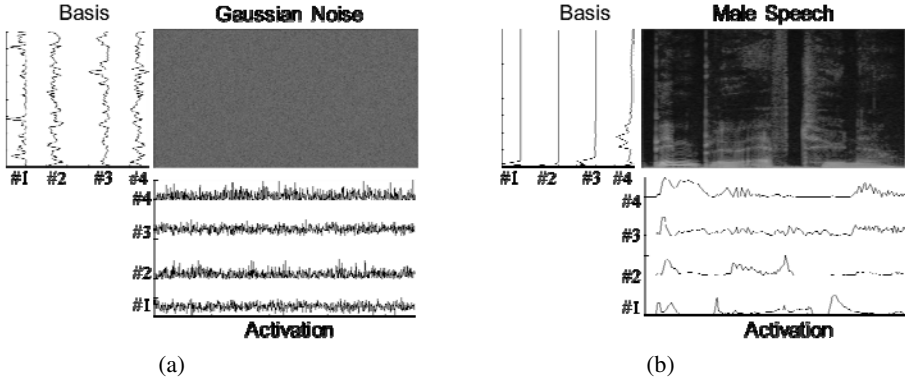

Fig. 2. Examples of NMF bases and activations for (a) Gaussian noise and (b) male speech

distribution for Gaussian noise and male speech, respectively. As shown in the figure, the spectrum of Gaussian noise is distributed across a wide range of frequencies and overlapped with that of male speech. Similarly, NMF bases of Gaussian noise are also widely distributed and overlapped with those of the male speech. Compared to the male speech, Gaussian noise, which is one of the typical stationary noises, can be removed well by conventional noise reduction methods such as a Wiener filter and an MMSE-LSA. Based on this observation, Wiener filtering and the NMF technique are combined in the proposed method.

Accordingly, it is assumed that $Y_k(\ell)$ is decomposed into stationary noise, $V_k(\ell)$, and non-stationary noise, $D_k(\ell)$; i.e., $Y_k(\ell) = V_k(\ell) + D_k(\ell)$. Assuming that a weight value, $G_{V,k}(\ell)$, for reducing stationary noise gives little damage on target speech, multiplying $G_{V,k}(\ell)$ to $X_k(\ell)$ provides the sum of the estimate of target speech and non-stationary noise such as $X_k(\ell) \cdot G_{V,k}(\ell) = \breve{S}_k(\ell) = \hat{S}_k(\ell) + D_k(\ell)$. As a next step, a weighting value, $G_{D,k}(\ell)$, for the residual noise attenuation is applied to $\breve{S}_k(\ell)$ in order to reduce non-stationary noise from $\breve{S}_k(\ell)$, which results in more enhanced target speech, $\hat{S}_k(\ell)$. That is, $\breve{S}_k(\ell) G_{D,k}(\ell) = \hat{S}_k(\ell)$. By combining these two steps, a weighting value, $G_k(\ell)$, for both the stationary and non-stationary noise reduction can be represented as the product of two weighting values, $G_{V,k}(\ell)$ and $G_{D,k}(\ell)$, such that $\hat{S}_k(\ell) = G_k(\ell) X_k(\ell)$, where $G_k(\ell) = G_{V,k}(\ell) \cdot G_{D,k}(\ell)$.

2.1 Stationary Noise Reduction Based on Wiener Filtering

In this subsection, we explain how to obtain $G_{V,k}(\ell)$ for stationary noise reduction. First, spectral variance of stationary noise, $\hat{\lambda}_{V,k}(\ell)$, is estimated by the recursive averaging method that is executed only when target speech absence is declared [11]. That is, $\hat{\lambda}_{V,k}(\ell) = \zeta_V \hat{\lambda}_{V,k}(\ell - 1) + (1 - \zeta_V) |X_k(\ell)|^2$ if target speech is absent, where ζ_V is a forgetting factor. Then, $G_{V,k}(\ell)$ is represented by employing the *a priori* SNR estimate by the decision-directed (DD) approach [3], $\hat{\xi}_k(\ell)$, as

$$G_{V,k}(\ell) = \frac{\hat{\xi}_k(\ell)}{\hat{\xi}_k(\ell) + 1}. \tag{1}$$

2.2 Non-stationary Noise Reduction Based on NMF

In this subsection, we explain how to estimate $G_{D,k}(\ell)$ for non-stationary (or residual) noise reduction by using the NMF technique. NMF is an algorithm for multivariate data analysis that decomposes a $K \times L$ matrix, $\mathbf{V}_{K \times L}$, into the product of a basis matrix, $\mathbf{B}_{K \times R}$, and an activation matrix, $\mathbf{A}_{R \times L}$; i.e., $\mathbf{V}_{K \times L} \approx \mathbf{B}_{K \times R} \mathbf{A}_{R \times L}$, where K, R, and L correspond to the number of spectral channels, the rank of the basis vector, and the number of frames, respectively. From now on, each matrix is

represented in the text without any subscript for the simplicity. To find $\mathbf{B}$ and $\mathbf{A}$, two kinds of cost functions are commonly used [6]: the Euclidean distance and the Kullback–Leibler (KL) divergence. For speech processing, NMF using the KL divergence shows better performance than that using the Euclidean distance [7], thus the KL divergence, $Div(\mathbf{X} \| \mathbf{BA})$, is used in this paper, defined as [12]

$$Div(\mathbf{X} \| \mathbf{BA}) = \sum_{i,j} \left(\mathbf{X}_{i,j} \log \frac{\mathbf{X}_{i,j}}{(\mathbf{BA})_{i,j}} - \mathbf{X}_{i,j} + (\mathbf{BA})_{i,j} \right) \tag{2}$$

where i and j indicate the row and column index of a matrix, respectively. By applying NMF, target speech estimated from Wiener filtering in Section 2.1, $\breve{\mathbf{S}}$, is further decomposed into target speech, $\hat{\mathbf{S}}$, and non-stationary noise, $\mathbf{D}$, by $Div(\breve{\mathbf{S}} \| \mathbf{BA})$, as

$$\breve{\mathbf{S}}_{K \times L} = \hat{\mathbf{S}}_{K \times L} + \mathbf{D}_{K \times L} \approx \left[\mathbf{B}_{S,K \times R_S}; \mathbf{B}_{D,K \times R_D} \right] \left[\mathbf{A}_{S,R_S \times L}; \mathbf{A}_{D,R_D \times L} \right] = \mathbf{B}_{K \times R} \mathbf{A}_{R \times L} \tag{3}$$

where R_S and R_D are the rank of the basis vectors for speech and non-stationary noise, respectively, and $R = R_S + R_D$. In Eq. 3, the basis matrix $\mathbf{B}(=[\mathbf{B}_S, \mathbf{B}_D])$ is replaced with the pre-trained matrix, $\overline{\mathbf{B}}(=[\overline{\mathbf{B}}_S, \overline{\mathbf{B}}_D])$, assuming that $\overline{\mathbf{B}}_S$ and $\overline{\mathbf{B}}_D$ hold the ability for constructing current speech and noise, respectively. Thus, we have $\mathbf{X} \approx \overline{\mathbf{B}}\mathbf{A}$. To obtain the non-stationary noise basis matrix, $\overline{\mathbf{B}}_D$, the non-stationary noise DB, $\overline{\mathbf{D}}$, is generated from the original noise DB, $\overline{\mathbf{Y}}$. As mentioned earlier, original noise is decomposed into non-stationary noise and stationary noise. Thus, the estimate of the variance, $\overline{\lambda}_{V,k}(\ell)$, is obtained by

$$\overline{D}_k(\ell) = \max(\overline{Y}_k(\ell) - \overline{\lambda}_{V,k}(\ell), 0)_{\forall k,\ell} \tag{4}$$

where $\overline{\lambda}_{V,k}(\ell)$, is the estimate of non-stationary noise by the recursive averaging method. To estimate the activation matrix, $\mathbf{A}(=[\mathbf{A}_S ; \mathbf{A}_D])$, the activation matrix $\mathbf{A}$ is first randomly initialized. Then, the cost function in Eq. 2 is minimized by iteratively applying an updating rule defined as [12]

$$\mathbf{A}^{m+1} = \mathbf{A}^m \otimes \frac{\mathbf{B}^{\mathrm{T}} \dfrac{\mathbf{X}}{\mathbf{BA}}}{\mathbf{B}^{\mathrm{T}}\mathbf{1}} \tag{5}$$

where m represents an iteration number, and $\mathbf{1}$ is a matrix with all elements equal to unity. Moreover, both multiplication $\otimes$ and division denote the element-wise operators. Hence, the weighting value for non-stationary noise reduction is represented as

$$\mathbf{G_D} = \frac{\overline{\mathbf{B}}_\mathbf{S}\mathbf{A_S}}{\overline{\mathbf{B}}_\mathbf{S}\mathbf{A_S} + \overline{\mathbf{B}}_\mathbf{D}\mathbf{A_D}} \tag{6}$$

or

$$G_{D,k}(\ell) = \frac{\sum_{rs=1}^{R_S} \left[\overline{B}_{S,k^i,rs} A_{S,rs,\ell^j} \right]}{\sum_{rs=1}^{R_S} \left[\overline{B}_{S,k^i,rs} A_{S,rs,\ell^j} \right] + \sum_{rd=1}^{R_D} \left[\overline{B}_{D,k^i,rd} A_{D,rd,\ell^j} \right]} \tag{7}$$

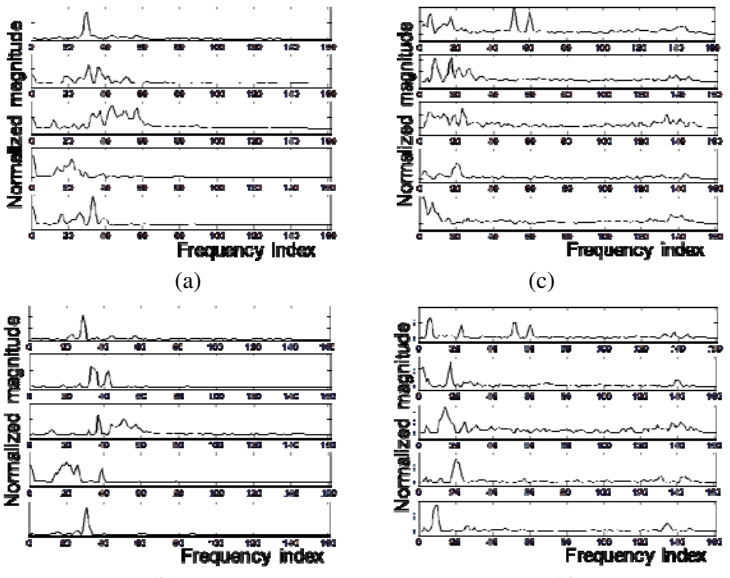

Fig. 3. Examples of noise bases for two difference noise signals ((a) and (c)) and the estimate of non-stationary noise ((b) and (d))

where k^i and ℓ^j indicate the row and column index of a matrix, respectively, and they correspond to the *k*-th frequency channel and the ℓ-th frame. Eq. 7 implies that the NMF-based noise reduction can also be interpreted as filtering the noisy signal with a time-varying filter, which is similar to the Wiener filtering in Eq. 2.

Fig. 3 shows the bases of the two different noise signals and the estimated non-stationary noise signals obtained from Eq. 7. It is shown from the figure that the bases of the estimated non-stationary noise (Figs. 3(b) and 3(d)) are more localized over frequency than those of the original noise bases (Figs. 3(a) and 3(c)).

2.3 Target Speech Reconstruction

The combined weighting value, $G_k(\ell)$, for noise reduction is represented as the product of $G_{V,k}(\ell)$ and $G_{D,k}(\ell)$ that are described in Eqs. 1 and 7, respectively. That is, $G_k(\ell) = G_{V,k}(\ell) G_{D,k}(\ell)$. Thus, the target speech estimate, $\hat{S}_k(\ell)$, is obtained by multiplying $G_k(\ell)$ to $X_k(\ell)$, and it is transformed into a time-discrete signal, $\hat{s}(n)$, that is finally brought to the MFCC extraction for speech recognition.

3 Speech Recognition Experiments

The performance of the proposed noise reduction method was evaluated in a view of ASR performance. First of all, a word recognition system in several different background noise environments was constructed, where acoustic models were

Table 1. Comparison of average word error rates (WERs) (%)

SNR	Bus Stop					Home TV				
(dB)	No	[3]	[4]	[8]	Proposed	No	[3]	[4]	[8]	Proposed
20	13.0	10.7	10.0	10.7	10.6	19.3	32.3	25.6	18.3	16.1
15	20.5	13.9	12.8	13.6	10.3	28.7	45.5	34.1	26.1	22.2
10	45.4	21.3	20.2	20.6	18.4	42.1	60.1	46.0	36.0	35.2
5	81.4	34.3	32.3	33.5	30.2	65.7	82.0	64.6	51.1	54.0
0	97.9	77.3	74.4	67.3	71.6	88.1	101.4	86.0	77.6	75.8
Avg.	51.6	31.5	30.0	29.1	28.2	48.8	64.3	51.3	41.8	40.7

SNR	Restaurant					Subway				
(dB)	No	[3]	[4]	[8]	Proposed	No	[3]	[4]	[8]	Proposed
20	14.6	11.5	12.7	12.8	11.8	11.9	13.4	12.7	12.9	10.5
15	22.6	17.7	15.3	15.1	15.2	16.3	13.2	13.2	13.2	10.2
10	48.7	24.2	22.2	21.6	20.1	35.1	19.2	16.8	18.8	14.3
5	86.2	43.0	36.1	39.2	34.8	74.5	33.6	31.6	33.3	28.4
0	100	80.8	72.1	70.1	71.4	97.5	67.7	69.2	66.0	64.0
Avg.	54.4	35.5	31.7	31.7	30.7	47.0	29.4	28.7	28.8	25.5

tri-phone based three-state left-to-right hidden Markov models (HMMs). The context-dependent acoustic models were trained from around 170,000 phonetically balanced words [5] recorded from 1,800 persons in quiet environments, where speech signals were sampled at a rate of 16 kHz with 16-bit resolution. As a speech recognition feature, a feature extraction procedure was applied once every 20 ms frame. In other words, 13 mel-frequency cepstral coefficients (MFCCs) including the zeroth order were extracted, and their first two derivatives were added, which resulted in a 39-dimensional feature vector per 20 ms frame.

A speech database was collected using a mobile phone, where there were 20 speakers (10 males and 10 females) and each speaker pronounced 40 utterances in a quiet office. On one hand, two sets of four different environmental noises were recorded such as bus stops, home TV, restaurants, and subways. A noise basis matrix was trained for each noise, whose length was 10 seconds long, from the first noise set. In order to obtain the NMF bases for speech, half of the speech database was used. Note here that each speaker had his/her own NMF bases that were kind of speaker-dependent NMF bases. In this paper, the rank of each basis vector for speech and noise were set at R_S=100 and R_D=50, respectively.

The other noise set was used to generate a test database. That is, each of half of utterances for a speaker was artificially added by each of four different environmental noises, where signal-to-noise ratios (SNRs) varied from 0 to 20 dB with a step of 5 dB. In total, there were 400 noisy speech utterances for the test.

Table 1 compares average word error rates (WERs) of an ASR system employing the proposed method with those employing conventional noise reduction methods

such as MMSE-LSA [3], the two-stage mel-warped Wiener filter (Mel-WF) [4], and the NMF-Wiener filter (NMF-WF) [8]. As shown in the table, MMSE-LSA gave the lowest performance in all noise environments under all SNR conditions. On the other hand, the Mel-WF and NMF-WF achieved similar WERs at bus stops, in restaurants, and on subways. However, in the home TV noise environment, NMF-WF outperformed Mel-WF. Note that the non-stationary components in home TV noise environment were more dominant than those in other noise environments. Comparing to NMF-WF, the proposed method provided smaller WER under all different SNRs and noise types. In other words, an ASR system employing the proposed method relatively reduced average WER by 5.2% compared to that using NMF-WF. Moreover, the proposed method provided WER reduction of 11.9% and 22.4% compared to Mel-WF and MMSE-LSA, respectively.

4 Conclusion

In this paper, we proposed an NMF-based noise reduction method for noise-robust ASR. To this end, stationary components in observed noisy speech were reduced by Wiener filtering. Next, an NMF-based decomposition technique was applied to remove the residual non-stationary noise that remained after the Wiener filter processing. In particular, the NMF bases of the residual noise were trained using the non-stationary noise database, estimated from an original noise database. It was shown from the ASR experiments that an ASR system employing the proposed method performed better than those using the conventional two-stage mel-warped Wiener filter, the MMSE-LSA estimator, and the NMF-Wiener filter.

References

1. Wu, J., Droppo, J., Deng, L., Acero, A.: A noise-robust ASR front-end using Wiener filter constructed from MMSE estimation of clean speech and noise. In: IEEE Workshop on ASRU, pp. 321–326 (2003)
2. Choi, H.C.: Noise robust front-end for ASR using spectral subtraction, spectral flooring and cumulative distribution mapping. In: 10th Australian Int. Conf. on Speech Science and Technology, pp. 451–456 (2004)
3. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33(2), 443–445 (1985)
4. Agarwal, A., Cheng, Y.M.: Two-stage mel-warped Wiener filter for robust speech recognition. In: IEEE Workshop on ASRU, pp. 67–70 (1999)
5. Lee, S.J., Kang, B.O., Jung, H.Y., Lee, Y.K., Kim, H.S.: Statistical model-based noise reduction approach for car interior applications to speech recognition. ETRI Journal 32(5), 801–809 (2010)
6. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
7. Wilson, K.K., Raj, B., Smaragdis, R., Divakaran, A.: Speech denoising using non-negative matrix factorization with priors. In: ICASSP, pp. 4029–4032 (2008)

8. Kim, S.M., Kim, H.K., Lee, S.J., Lee, Y.K.: Noise robust speech recognition based on a non-negative matrix factorization. In: Inter-noise 2011 (2011)
9. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Speech Audio Process. 15(3), 1066–1074 (2007)
10. Malah, D., Cox, R., Accardi, A.J.: Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments. In: ICASSP, pp. 789–792 (1999)
11. Sohn, J., Kim, N.S., Sung, W.: statistical model-based voice activity detection. IEEE Signal Process. Lett. 6(1), 1–3 (1999)
12. Lee, D.D., Seung, H.S.: Algorithms for nonnegative matrix factorization. In: Adv. Neural Inform. Process. Sys., vol. 13, pp. 556–562 (2000)