SPEECH ENHANCEMENT COMBINING STATISTICAL MODELS AND NMF WITH UPDATE OF SPEECH AND NOISE BASES

Kisoo Kwon[⋆]*, Jong Won Shin*[†]*, Sukanya Sonowal*[⋆]*, Inkyu Choi*[⋆] *and Nam Soo Kim*[⋆]

[⋆]Department of Electrical and Computer Engineering and INMC
Seoul National University, Seoul 151-742, Korea
[†]School of Information and Communications
Gwangju Institute of Science and Technology, Gwangju, Korea
E-mail: kskwon@hi.snu.ac.kr, jwshin@gist.ac.kr, sukanya@hi.snu.ac.kr, ikchoi@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

Speech enhancement based on statistical models has shown good performance, but the performance degrades when environment noise is highly non-stationary due to the stationary assumption. On the contrary, the template-based enhancement methods are more robust to non-stationary noise, but these are heavily dependent on *a priori* information present in training data. In order to get over both of the shortcomings, we propose a novel speech enhancement method which combines the statistical model-based enhancement scheme with the template-based enhancement. To reduce a dependency on *a priori* information, the speech and noise bases are updated simultaneously using the estimated speech presence probability, which is obtained from statistical model-based enhancement. Experimental results showed that the proposed method outperformed not only the statistical model-based and non-negative matrix factorization (NMF) approaches, but also their combination implemented with existing bases update rule in various kinds of noise.

Index Terms— Statistical model-based enhancement, non-negative matrix factorization, on-line update of bases.

1. INTRODUCTION

Speech enhancement has been extensively studied in the last few decades, and various approaches have been proposed. Single channel speech enhancement techniques can be broadly classified into the statistical model-based and template-based approaches [1]-[9]. In the statistical model based approach, speech and noise are modelled with separate parametric distributions for which each parameter is estimated from the input signal [1]-[3]. One of the important advantages of the statistical model-based techniques is that the models do not need to be trained *a priori*. However, since the statistical models are made based on the stationarity assumption of speech and noise, the performance deteriorates when the environment noise is highly non-stationary.

On the other hand, the template-based techniques, unlike the statistical model-based methods, need the *a priori* information of speech or noise [4]-[6]. *A priori* information can be statistics obtained from a speech or noise corpus or typical patterns [4], [5]. These approaches are more robust to non-stationary environments, since there is no strict assumption on the nature of the noise in contrast to the statistical model-based methods. However, if the noise is far different from the trained noise model, the performance degrades severely [6]. Therefore, in order to show good performance they require a sufficiently large and rich training data set of various noise environments.

A number of methods have been proposed to combine the aforementioned two techniques to attain better performance. A template-based method estimate the speech magnitude spectrum in [4], while it is applied to obtain the noise power spectral density (PSD) in [7]. These two methods compute Wiener filter type gain functions using the PSDs derived from the template-based approaches. In contrast, [8] applies a template-based algorithm at the output of a Wiener filter. This method takes advantage of both the statistical model-based and template-based approaches, but the Wiener filter output may not be modelled well without any bases update. In [9], non-negative matrix factorization (NMF)-based enhancement is combined with voice activity detection (VAD) which is obtained by statistical models, but the performance degrades if the trained noise model is different from the actual noise environment.

In this paper, a cascaded structure that combines the

This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A01045874) and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4005) supervised by the NIPA(National IT Industry Promotion Agency) and Basic Science Research Program through the NRF (NRF-2013R1A1A1013418).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

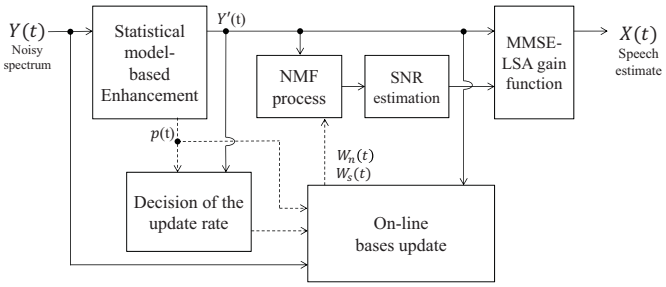

Fig. 1. Block diagram of the proposed speech enhancement method

statistical model-based enhancement and template-based approaches along with simultaneous update of speech and noise bases. The bases are updated using the speech presence probability (SPP), and so the proposed method can deal with the speech and noise patterns which were not included in the training database well. Experimental results showed that the proposed algorithm outperformed not only the statistical model-based and NMF-based methods but also the combined approach with conventional bases update rule.

2. PROPOSED SPEECH ENHANCEMENT METHOD WITH ON-LINE BASES UPDATE

In this paper, we propose a cascaded structure for speech enhancement of which the first stage is a statistical model-based enhancement while the second stage consists of NMF-based noise reduction with an on-line update of both speech and noise bases. The overall block diagram is illustrated in Figure1. The first stage performs statistical model-based enhancement (SE), which produces pre-enhanced signal and SPP for the second stage and bases update module. The second stage implements an NMF-based enhancement module where the minimum mean square error-log spectral amplitude (MMSE-LSA) estimator [10] is adopted for which the signal-to-noise-ratio (SNR)-related parameters are estimated by the NMF module. In our approach, both the speech and noise bases are updated with the help of SPP to deal with noise unseen during training and to correct the SE output which is not well-covered by the original bases.

2.1. NMF-based enhancement and MMSE-LSA gain function

The second stage of enhancement adopts NMF analysis based on the algorithm presented in [11], [12] with Kullback-Leibler divergence as the distance metric. A data set $V \in \mathbb{R}^{M\times N}$ is reconstructed by the product of basis matrix $W \in \mathbb{R}^{M\times r}$ and the encoding matrix $H \in \mathbb{R}^{r\times N}$ ($V \approx WH$), where M and N respectively denote the number of frequency bins and time frames, and r is the number of bases. We can consider that W consists of speech bases $W_s \in \mathbb{R}^{M\times r_s}$ and noise bases $W_n \in \mathbb{R}^{M\times r_n}$, i.e., $W = [W_s\ W_n] \in \mathbb{R}^{M\times(r_s+r_n)}$, where r_s and r_n indicate the numbers of corresponding bases. To apply the NMF framework to an audio spectrum, $V(t)$ which is the *t*-th frame of input, is constructed as $V(t) = |Y'(t)|$ where $|\cdot|$ means taking element-wise magnitude, and $Y'(t) \in \mathbb{C}^{M\times 1}$ is SE output spectrum while $Y(t) \in \mathbb{C}^{M\times 1}$ denote the input spectrum. Since $Y'(t)$ would have better SNR, the input to the second stage of enhancement is SE output.

Let $W_s(t)$ and $W_n(t)$ denote the updated speech and noise base which will be used to analyze the *t*-th frame of pre-enhanced signal, i.e., $W(t) = [W_s(t)\ W_n(t)]$. Given $W(t)$ which is updated by methods in the next subsections, $H(t) = [H_s(t)^T\ H_n(t)^T]^T \in \mathbb{R}^{(r_s+r_n)\times 1}$ is optimized for fixed basis $W(t)$ and $V(t)$ by multiplicative method [11], [12] with ^{T} denoting matrix transpose.

$$H(t) \leftarrow H(t) \otimes \frac{W^T(t)\frac{V(t)}{W(t)H(t)}}{W^T(t)\mathbf{1}}. \tag{1}$$

where $\bigotimes$ and $\frac{a}{b}$ are the element-wise multiplication and division of matrices, and $\mathbf{1} \in \mathbb{R}^{M\times 1}$ is the all-one vector. $H_s(t)$ is initialized by $H_s(t-1)$ and $H_n(t)$ is initialized by random values for every frame, and (1) is repeated until the convergence.

Then, the speech and noise magnitude spectra estimates, $\hat{S}(t)$ and $\hat{N}(t)$, can be obtained as

$$\hat{S}(t) = W_s(t)H_s(t), \qquad \hat{N}(t) = W_n(t)H_n(t). \tag{2}$$

Using the speech and noise magnitude spectra estimates, the spectral gain function $G(m,t), m = 1, \cdots, M$, is computed. In contrast to the approaches in [5] and [8] which used Wiener filter, the MMSE-LSA [10] is adopted in this paper, which leads to the gain function.

$$G(m,t) = \frac{\xi(m,t)}{1+\xi(m,t)} \exp\left(\frac{1}{2}\int_{\nu(m,t)}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{3}$$
$$\nu(m,t) = \frac{\gamma(m,t)\xi(m,t)}{1+\xi(m,t)}$$

in which $\xi(m,t)$ is the a priori SNR and $\gamma(m,t)$ is the a posteriori SNR for the *m*-th frequency bin at frame *t*. They are estimated as follows using smoothing factors τ_s and τ_n

$$P_s(m,t) = \tau_s P_s(m,t-1) + (1-\tau_s)[(\hat{S}(t))_m]^2 \tag{4}$$
$$P_n(m,t) = \tau_n P_n(m,t-1) + (1-\tau_n)[(\hat{N}(t))_m]^2$$
$$\xi(m,t) = \frac{P_s(m,t)}{P_n(m,t)} \quad \gamma(m,t) = \frac{[(V(t))_m]^2}{P_n(m,t)}$$

where $P_s(m,t)$ and $P_n(m,t)$ denote the smoothed speech and noise PSDs for the *m*-th frequency bin at frame *t*, respectively. Finally, the enhanced speech spectrum is calculated according to $\hat{X}(m,t) = G(m,t)Y'(m,t)$ when $Y'(m,t)$ means the *m*-th element of $Y'(t)$.

2.2. On-line update of speech and noise bases

In NMF bases update, both the noisy input spectrum $Y(t)$ and the SE output $Y'(t)$ can be the candidates of the input for NMF analysis. $Y(t)$ contains complete information on original speech and noise but the SNR is low. On the contrary, $Y'(t)$ would have better SNR than $Y(t)$ but may have already lost some of the weak speech components. Based on this observation, the input to NMF for updating the bases is constructed with both of the spectra, i.e., $\tilde{V}(t) = [|Y'(t)| \; |Y(t)|] \in \mathbb{R}^{M\times 2}$. The optimization method is the same to [11], [12] and the update at each iteration is performed by

$$\tilde{H}(t) \leftarrow \tilde{H}(t) \otimes \frac{\tilde{W}^T(t)\frac{\tilde{V}(t)}{\tilde{W}(t)\tilde{H}(t)}}{\tilde{W}^T(t)\tilde{\mathbf{1}}}, \tag{5}$$

$$\tilde{W}(t) \leftarrow \tilde{W}(t) \otimes \frac{\frac{\tilde{V}(t)}{\tilde{W}(t)\tilde{H}(t)}\tilde{H}^T(t)}{\tilde{\mathbf{1}}\tilde{H}^T(t)}.$$

where $\tilde{\mathbf{1}} \in \mathbb{R}^{M\times 2}$ is all-one matrix, and $\tilde{H}(t) \in \mathbb{R}^{(r_s+r_n)\times 2}$ and $\tilde{W}(t) = [\tilde{W}_s(t) \; \tilde{W}_n(t)]$ are the instantaneous encoding and basis matrices of *t*-th frame, respectively. In our work, $\tilde{H}(t)$ is initialized with random values and $\tilde{W}(t)$ is initialized by $W(t-1)$ at each time frame. The iteration of (5) is performed continuously until convergence, which does not require too many iterations.

The instantaneous speech and noise bases matrix $\tilde{W}(t)$ obtained through (5) usually emphasizes the current input $Y(t)$ and $Y'(t)$, resulting in an abrupt change of the bases. Also, when the bases are updated, it is very important to discriminate the speech and noise components correctly for speech enhancement.

To resolve this problem, we use the SPP $\boldsymbol{p}(t) \in \mathbb{R}^{M\times 1}$ each element of which indicates the probability of speech presence in a corresponding frequency bin, to control the speech and noise bases update. This SPP can be estimated in the SE stage. The speech and noise bases are updated given as follows:

$$W_s(t) = \boldsymbol{\lambda}_s(t) \otimes \tilde{W}_s(t) + (\mathbf{1}_{M\times r_s} - \boldsymbol{\lambda}_s(t)) \otimes W_s(t-1), \tag{6}$$

$$\boldsymbol{\lambda}_s(t) = \alpha_s \boldsymbol{p}(t)\mathbf{1}_{r_s},$$

$$W_n(t) = \boldsymbol{\lambda}_n(t) \otimes \tilde{W}_n(t) + (\mathbf{1}_{M\times r_n} - \boldsymbol{\lambda}_n(t)) \otimes W_n(t-1), \tag{7}$$

$$\boldsymbol{\lambda}_n(t) = \alpha_n(\mathbf{1}_{M\times r_n} - \boldsymbol{p}(t)\mathbf{1}_{r_n}).$$

where α_s and α_n are the maximum update rates for $\tilde{W}_s$ and $\tilde{W}_n$, and $\mathbf{1}_{M\times r_s} \in \mathbb{R}^{M\times r_s}$, $\mathbf{1}_{M\times r_n} \in \mathbb{R}^{M\times r_n}$, $\mathbf{1}_{r_s} \in \mathbb{R}^{1\times r_s}$ and $\mathbf{1}_{r_n} \in \mathbb{R}^{1\times r_n}$ are all-one matrices. In (6) and (7), $\boldsymbol{\lambda}_s(t) \in \mathbb{R}^{M\times r_s}$ and $\boldsymbol{\lambda}_n(t) \in \mathbb{R}^{M\times r_n}$ determine the rates of bases update depending on $\boldsymbol{p}(t)$. This update rule enables a robust update of the speech and noise bases leading to a stable speech enhancement performance. α_s and α_n are experimentally determined.

Table 1. PESQ score improvement for various noises with matched noise bases. (input SNR: 5 dB)

Noise Type	*Leo.*	*F-16*	*bucc.*	*hfch.*	Average
SE	0.3018	0.5116	0.5349	0.5422	0.4726
NMF	0.4854	0.2091	0.7321	0.8768	0.5759
SE & Cabras	0.5247	0.5402	0.6263	0.7081	0.5998
Proposed w/o update	0.7729	0.6579	0.7814	0.9661	0.7946
Proposed	0.8370	0.6984	0.8079	1.0220	**0.8413**

3. EXPERIMENT

In this paper, among a number of statistical model-based enhancements, we adopted the algorithm presented in [13] for the SE stage because of its simplicity and good performance. This algorithm applies Winer filter type gain function, and produces pre-enhanced signal and the SPP.

Speech and noise audio samples were obtained from TIMIT and NOISEX-92 databases (DBs), respectively, and the sampling rate was 16 kHz. A 75% overlap was used along with 512 point fast Fourier transform. Each noise basis was trained using 16 s-long noise signal which was not included in the test DB. The number of speech and noise bases were 40 each ($r_n = r_s = 40$). The parameter values related to smoothing were $\tau_s = 0.5$ and $\tau_n = 0.9$. The update rate for speech bases was fixed at $\alpha_s = 0.03$, and the update rate for noise bases α_n was fixed to 0.03 for the first and the third experiments, while it was fixed to 0.1 for the second test.

The performance measure was the ITU-T Recommendation P.862 Perceptual evaluation of speech quality (PESQ) [14] score. If enhanced signal is close to clean speech, then PESQ score is 4.5, while least score is 0.

The performance of the following five systems were compared:

◦ *SE* : Only SE [13] was applied.

◦ *NMF* : Only NMF based-enhancement was used without bases update.

◦ *SE & Cabras* : The NMF-based enhancement proposed by Cabras et al. [9], which updates speech and noise bases based on VAD was applied. When speech is absent, only the noise basis is updated, while only the speech basis is updated in the presence of speech signal. For the fair comparison, MMSE-LSA is adopted as the spectral gain function, and pre-enhanced signal from SE is used for the input, just like our proposed method.

◦ *Proposed w/o update* : A cascade form that combines *SE* and *NMF* without the bases update.

◦ *Proposed* : A cascade form with the on-line bases update.

Table 2. PESQ score improvement for various noises for which noise bases were trained with white noise.(input SNR : 5 dB)

Noise Type	*Leo.*	*F-16*	*bucc.*	*hfch.*	Average
SE	0.3018	0.5116	0.5349	0.5422	0.4726
NMF	-0.0394	0.1304	0.1299	0.0008	0.0055
SE & Cabras	0.2538	0.5003	0.5190	0.5227	0.4490
Proposed w/o update	0.2530	0.5529	0.6115	0.5396	0.4893
Proposed	0.5958	0.6919	0.7577	0.9861	**0.7579**

Table 3. PESQ score improvement for various noises mixed with *machinegun* noise with matched noise bases.

Noise Type	*Leo.*	*F-16*	*bucc.*	*hfch.*	Average
SE	-0.0325	0.0739	0.1604	0.1819	0.0959
Proposed w/o update	0.8955	0.8567	1.0081	1.0961	0.9641
Proposed	0.8556	0.9263	1.0160	1.0967	**0.9737**

Three experiments were conducted to demonstrate the performances of the proposed method 1) in stationary noise with matched noise bases, 2) in the same noise with mismatched noise bases, and 3) in non-stationary noise with matched noise bases, respectively.

Matched noise bases were trained with the same kind of noise DB as the noise used for test. For example, *F-16 cockpit (F-16)* noise bases were trained by *F-16* noise DB. On the contrary, mismatched noise bases were made from noise DB different from the actual noise in the test signal, and *white* noise was used in this experiment. In the third experiment, only the noise bases were updated because the SPP from SE stage can not detect non-stationary noise reliably.

Table 1 shows the PESQ score improvement obtained with somewhat stationary noises such as *Leopard (Leo.)*, *F-16*, *buccaneer1 (bucc.)* and *hfchannel (hfch.)* noises when the tested noise type was included in the training DB. However, training DB did not include the noise excerpts used in the test. The SNR for each noise type was set to 5 dB. The result shows that the proposed method outperformed other enhancement systems. It is also clear that the on-line bases update was effective even when the trained W_n matched to the test DB.

Table 2 shows the PESQ score improvement obtained in the same noise environment as Table 1, but this time W_n was trained with *white* noise only. We can see that performance of the method with NMF only degrades significantly due to the heavy dependency on a priori information. It is clear that the update of bases made significant improvement on NMF performance.

The PESQ score improvement obtained when the non-stationary *machinegun* noise was present as well as the stationary noises used for the previous experiments are presented in Table 3. The noise level of *machinegun* was 5 dB lower than the speech level. It can be demonstrated that the proposed method could also deal with non-stationary noise well.

4. CONCLUSION

This paper has proposed a speech enhancement method which combines statistical model-based approach and the NMF approach with speech and noise bases update. Speech presence probability is applied to control the speech and noise bases update. By the combination of two distinct approaches and on-line speech and noise bases update, non-stationary noises can be handled and the dependency on the *a priori* information of speech and noise is reduced. Through the experiments, it was shown that the proposed algorithm performed better than the other methods regardless of noise stationarity or appropriateness of trained noise bases.

5. REFERENCES

[1] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Lett.*, vol. 7, no. 5, May 2000.

[2] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403-2418, Nov. 2001.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol.9, no.5, pp. 504-512. 2001.

[4] J. Ming, R. Srinivasan and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822-836, May 2011.

[5] K. W. Wilson, B. Raj, P. Smaragdis and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," *in IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[6] S. Srinivasan, J. Samuelsson and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transaction on Audio, Speech, And Language Processing*, vol. 14, no. 1, pp. 163-176, Jan. 2006.

[7] N. Mohammadiha, T. Gerkmann and A. Leijon "A new approach for speech enhancement based on a constrained NMF," *in Proc. IEEE Int. Symp. Intell. Signal Process. and Commun. Syst. (ISPACS)*, pp. 1-5, 2011.

[8] S. M. Kim, J. H. Park, H. K. Kim, S. J. Lee and Y. K. Lee, "Non-negative matrix factorization based noise reduction for noise robust automatic speech recognition," *Lecture Notes in Computer Science*, vol. 7191, pp. 338-346, 2012.

[9] G. Cabras, S. Canazza, P. L. Montessoro and R. Rinaldo, "Restoration of audio documents with low SNR: a NMF parameter estimation and perceptually motivated Bayesian suppression rule," *in Proc. Sound and Music Computing Conference*, pp. 314-321, 2010.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[12] H. S. Seung and D. D. Lee, "Algorithms for nonnegative matrix factorization," *Advances in neural information processing systems*, 13, pp. 556-562, 2001.

[13] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, pp. 220-231, 2006.

[14] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep. ITU-T P.862, 2001.