BASIS COMPENSATION IN NON-NEGATIVE MATRIX FACTORIZATION MODEL FOR SPEECH ENHANCEMENT

Hanwook Chung[1]*, Eric Plourde*[2] *and Benoit Champagne*[1]

[1]Dept. of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada
[2]Dept. of Electrical and Computer Engineering, Sherbrooke University, Sherbrooke, Quebec, Canada
e-mail:hanwook.chung@mail.mcgill.ca, eric.plourde@usherbrooke.ca, benoit.champagne@mcgill.ca

ABSTRACT

In this paper, we propose a basis compensation algorithm for non-negative matrix factorization (NMF) models as applied to supervised single-channel speech enhancement. In the proposed framework, we use extra free basis vectors for both the clean speech and noise during the enhancement stage in order to capture the features which are not included in the training data. Specifically, the free basis vectors of the clean speech are obtained by exploiting *a priori* knowledge based on a Gamma distribution. The free bases of the noise are estimated using a regularization approach, which enforces them to be orthogonal to the clean speech and noise basis vectors estimated during the training stage. Experimental results show that the proposed NMF algorithm with basis compensation provides better performance in speech enhancement than the benchmark algorithms.

Index Terms— Single-channel speech enhancement, non-negative matrix factorization, supervised algorithm, basis adaptation

1. INTRODUCTION

Numerous algorithms for single-channel speech enhancement have been proposed in the past such as: minimum mean-square error (MMSE) estimation [1], spectral subtraction [2], and subspace decomposition [3]. These algorithms, however, use a minimal amount of *a priori* information about the speech and noise and hence, tend to provide limited performance under adverse noise conditions. Recently, the non-negative matrix factorization (NMF) approach, which decomposes a given matrix into basis and activation matrices with non-negative element constraints [4, 5], has been successfully applied to diverse problems such as image representation [6], source separation [7] and speech enhancement [8]. In speech and audio applications, the magnitude or power spectrum is interpreted as a linear combination of the basis vectors which can be obtained *a priori* using training data.

In a supervised NMF-based framework, the basis vectors are obtained for each source independently during the training stage, and used subsequently in the separation or enhancement stage. One main problem of such supervised algorithms is the existence of a mismatch between the characteristics of the training and test data which in turn leads to a decreased quality of the estimated sources. A possible remedy to this problem is to add explicit regularization terms to the NMF cost function that incorporate some prior knowledge, such as temporal continuity [9] or statistical priors of the magnitude spectra [10]. In these algorithms, however, the basis vectors are fixed during the enhancement stage, which limits the performance when there is a large mismatch between the training and test data.

One alternative approach to handle such problem is to use a basis adaptation scheme during the enhancement stage. In [11], the basis vectors are adapted based on prior distributions modeled by Gamma mixtures. In [12], a basis adaptation scheme for sparse convolutive NMF model has been proposed. The authors in [13] employ extra validation data for speaker adaptation in a speech-music separation task. Recently in [14], the basis vectors are adapted by using a combination of the original and pre-processed noisy speech samples, the latter being obtained via a classical MMSE-based speech enhancement algorithm. The estimated basis vectors are further corrected based on the speech presence probability (SPP). In these algorithms, however, the basis vectors are adapted from the mixtures of multiple sources, e.g., noisy speech, such that the resulting basis vectors may still exhibit features of different sources. Consequently, the enhanced speech may contain some residual noise components. Hence, adapting the complete set of basis vectors may limit the enhanced speech quality.

In this paper, we propose a new basis adaptation algorithm for NMF-based speech enhancement, motivated by semi-supervised applications where the training data are available for only a few sources [15]. In the proposed framework, we use extra free basis vectors for both the clean speech and noise during the enhancement stage in order to capture the features which are not included in the training data. Specifically, the free basis vectors of the clean speech are obtained by exploiting *a priori* information about the basis vectors based on the Gamma distribution [16]. The free basis vectors of the noise are estimated using a regularization approach, which enforces them to be orthogonal to the clean speech and noise basis vectors estimated during the training stage. [15, 17]. These free basis vectors are estimated from the noisy speech along with the pre-processed signals similar to [14]. Since we use extra free basis vectors and do not change the trained ones, we refer to the proposed method as basis *compensation* rather than *adaptation*. Experimental results of perceptual evaluation of speech quality (PESQ) [23], source-to-distortion ratio (SDR) [24] and segmental SNR (SSNR) show that the proposed algorithm provides better enhancement performance than the benchmark algorithms.

In this paper, we use the subscripts or superscripts Y, S and N for indicating the noisy speech, clean speech and noise, respectively. We use the bold upper case to denote the matrix, e.g., $\mathbf{H}$, and bold lower case for the column vector, e.g., $\mathbf{y}$. The symbol $\mathbb{R}_+$ denotes the set of non-negative real numbers and $\mathbf{1}_{KL}$ is a $K \times L$ matrix with all entries equal to one.

Funding for this work was provided by Microsemi Corporation (Ottawa, Canada) and a grant from NSERC (Govt. of Canada).

2. NMF-BASED SPEECH ENHANCEMENT

For a given matrix $\mathbf{V} = [v_{kl}] \in \mathbb{R}_+^{K\times L}$, NMF finds a local optimal decomposition of $\mathbf{V} = \mathbf{WH}$, where $\mathbf{W} = [w_{km}] \in \mathbb{R}_+^{K\times M}$ is a basis matrix, $\mathbf{H} = [h_{ml}] \in \mathbb{R}_+^{M\times L}$ is an activation matrix and M is the number of basis vectors. The factorization is obtained by minimizing a cost function, such as the Kullback-Leibler (KL) divergence which is defined as,

$$\mathcal{D}_{\mathrm{KL}}(\mathbf{V},\mathbf{WH}) = \sum_k \sum_l \Big(v_{kl} \ln \frac{v_{kl}}{[\mathbf{WH}]_{kl}} - v_{kl} + [\mathbf{WH}]_{kl} \Big) \quad (1)$$

where $[\cdot]_{kl}$ denotes the (k,l)-th entry of its matrix argument. The solution can be obtained via multiplicative update rules [4]:

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\mathbf{V}/(\mathbf{WH}))\mathbf{H}^T}{\mathbf{1}_{KL}\mathbf{H}^T}, \qquad \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{V}/(\mathbf{WH}))}{\mathbf{W}^T\mathbf{1}_{KL}} \quad (2)$$

where the operation $\otimes$ denotes element-wise multiplication, / and the quotient line are element-wise division and the superscript T is the matrix transpose. After each iteration, the columns of $\mathbf{W}$ are normalized using the l_1-norm and the rows of $\mathbf{H}$ are scaled, accordingly, in order to avoid the scale indeterminacies [5]. As for the initializations of $\mathbf{W}$ and $\mathbf{H}$, positive random numbers are often used.

In NMF-based single-channel speech enhancement, we assume in practice that the magnitude spectrum of the noisy speech, obtained via short-time Fourier transform (STFT), can be approximated by the sum of the clean speech and noise magnitude spectra, i.e., $|\mathbf{y}_l| \approx |\mathbf{s}_l| + |\mathbf{n}_l|$, where $\mathbf{y}_l = [Y_{kl}]$, $\mathbf{s}_l = [S_{kl}]$ and $\mathbf{n}_l = [N_{kl}]$ respectively denote the noisy speech, clean speech and noise spectra, and $k \in \{1,\ldots,K\}$ and $l \in \{1,\ldots,L\}$ are the frequency bin and time frame indices [7, 8, 14]. In a supervised framework, $\mathbf{W}_S = [w^S_{km}] \in \mathbb{R}_+^{K\times M_S}$ and $\mathbf{W}_N = [w^N_{km}] \in \mathbb{R}_+^{K\times M_N}$ are obtained during the training stage, by applying (2) to the training data of the clean speech and noise magnitude spectra. In the enhancement stage, by fixing $\mathbf{W}_Y = [\mathbf{W}_S\ \mathbf{W}_N]$, the activation vector $\mathbf{h}^Y_l = [(\mathbf{h}^S_l)^T\ (\mathbf{h}^N_l)^T]^T \in \mathbb{R}_+^{(M_S+M_N)\times 1}$ is estimated at the l-th time frame by applying the activation update to $|\mathbf{y}_l|$.

The clean speech can be estimated from the noisy speech spectrum using a gain function, as in $\hat{S}_{kl} = G_{kl}Y_{kl}$. Among various choices for G_{kl}, we use the well-known MMSE short-time spectral amplitude (STSA) estimator [1],

$$G_{kl} = \Gamma(1.5)\frac{\sqrt{\rho_{kl}}}{\gamma_{kl}} \exp\Big(-\frac{\rho_{kl}}{2}\Big)\Big[(1+\rho_{kl})I_0\Big(\frac{\rho_{kl}}{2}\Big)+\rho_{kl}I_1\Big(\frac{\rho_{kl}}{2}\Big)\Big] \quad (3)$$

where $\Gamma(\cdot)$ is the Gamma function, and $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zero and first order, respectively. The quantity ρ_{kl} is defined as follows in terms of the *a priori* SNR, ξ_{kl}, and *a posteriori* SNR, γ_{kl}:

$$\rho_{kl} = \frac{\xi_{kl}}{1+\xi_{kl}}\gamma_{kl}, \quad \xi_{kl} = \frac{\hat{p}^S_{kl}}{\hat{p}^N_{kl}}, \quad \gamma_{kl} = \frac{|Y_{kl}|^2}{\hat{p}^N_{kl}} \quad (4)$$

where $\hat{p}^S_{k,l}$ and $\hat{p}^N_{k,l}$ respectively denote the estimated power spectral densities (PSD) of the clean speech and noise. The latter are obtained via temporal smoothing of the NMF-based periodograms as [14],

$$\hat{p}^S_{k,l} = \tau_S\hat{p}^S_{k,l-1} + (1-\tau_S)\Big(\sum_m w^S_{km}h^S_{ml}\Big)^2 \quad (5)$$

$$\hat{p}^N_{k,l} = \tau_N\hat{p}^N_{k,l-1} + (1-\tau_N)\Big(\sum_m w^N_{km}h^N_{ml}\Big)^2 \quad (6)$$

where τ_S and τ_N are the smoothing factors for the clean speech and noise, respectively. The time-domain enhanced speech signal is obtained via inverse STFT followed by the overlap-add method.

3. PROPOSED ALGORITHM

3.1. Training stage

During the training stage, we obtain basis matrices for the clean speech and noise, $\mathbf{W}_S$ and $\mathbf{W}_N$, by applying the update rules given in (2) to the corresponding training data, separately. In addition, we estimate an extra basis matrix for the clean speech, i.e., $\mathbf{W}_{Sp} = [w^{Sp}_{km}] \in \mathbb{R}_+^{K\times M_{SF}}$ such that $M_{SF} < M_S$, whose entries will be used as hyper-parameters in the *a priori* distribution of the free basis matrix of the clean speech, during the enhancement stage, as explained in the next section.

3.2. Enhancement stage

In the proposed framework, we use the classical MMSE STSA estimator [1] as a pre-processor, where the noise PSD is estimated based on [18]. This pre-processing removes some of the background noise and the NMF-based enhancement algorithm is applied subsequently to further improve the enhancement performance [14, 19]. Moreover, we can exploit different features from the ones obtained via NMF-based algorithms, which can be useful for the basis adaptation [14]. The proposed enhancement stage consists of two parts, first the free basis computation and, second, the actual enhancement, as explained in the next section.

3.2.1. Free basis computation

For the l-th time frame, the free basis matrices of the clean speech and noise, $\mathbf{W}^{SF}_l = [w^{SF}_{km,l}] \in \mathbb{R}_+^{K\times M_{SF}}$ and $\mathbf{W}^{NF}_l = [w^{NF}_{km,l}] \in \mathbb{R}_+^{K\times M_{NF}}$, are obtained. The goal is to estimate $\mathbf{W}^{SF}_l$ by exploiting *a priori* knowledge, whereas $\mathbf{W}^{NF}_l$ is obtained to capture the features which are not included in the training data of the clean speech or noise. Specifically, we consider the Gamma distribution for the prior of $\mathbf{W}^{SF}_l$, which is shown to be a conjugate prior to the NMF model with KL-divergence in a statistical framework [16]. The prior distribution for each entry of $\mathbf{W}^{SF}_l$ is given by,

$$p(w_{km}|\alpha_{km},\beta^{-1}_{km}) = (w_{km})^{\alpha_{km}-1}\beta^{\alpha_{km}}_{km}e^{-w_{km}\beta_{km}}/\Gamma(\alpha_{km}) \quad (7)$$

where we have omitted l and *SF* for notational convenience, α_{km} and β_{km} are the hyper-parameters, which in practice can be selected as $\alpha_{km} = 2$ and $\beta_{km} = (w^{Sp}_{km})^{-1}$ [13]. For $\mathbf{W}^{NF}_l$, we use a regularization approach, which enforces its column vectors to be orthogonal to every column vector in $[\mathbf{W}_S\ \mathbf{W}_N]$ [15, 17]. These free basis matrices are obtained by estimating 1) the instantaneous free basis matrices for the clean speech and noise, $\mathbf{W}_{SI} = [w^{SI}_{km}]$ and $\mathbf{W}_{NI} = [w^{NI}_{km}]$, followed by 2) some further corrections applied on these matrices. These two steps are detailed below.

1) Instantaneous basis computation: The noisy speech, $\mathbf{y}_l$, contains the complete information of the clean speech but also includes some noise components. On the contrary, less noise components are found in the pre-processed speech, which will be denoted as $\bar{\mathbf{s}}_l$, but some clean speech features are attenuated. Therefore, it is necessary to take into account both $\mathbf{y}_l$ and $\bar{\mathbf{s}}_l$ as the target values for the basis update [14]. We construct the target matrix by concatenating the observations as $\mathbf{V}_I = [\ |\bar{\mathbf{s}}_l|,\ |\bar{\mathbf{n}}_l|,\ |\mathbf{y}_l|\ ] = [v_{kq}] \in \mathbb{R}_+^{K\times 3}$ where $\bar{\mathbf{n}}_l = \mathbf{y}_l - \bar{\mathbf{s}}_l$. The instantaneous free basis matrix $\mathbf{W}_I = [\mathbf{W}_{SI}\ \mathbf{W}_{NI}] = [w^I_{km}]$ and activation matrix $\mathbf{H}_I = [h^I_{mq}] \in \mathbb{R}_+^{(M_{SF}+M_{NF})\times 3}$ are estimated from $\mathbf{V}_I$. The proposed cost function is shown as,

$$\mathcal{J} = \mathcal{D}_{\mathrm{KL}}(\mathbf{V}_I, \mathbf{W}_I\mathbf{H}_I) - \mathcal{R}_P(\mathbf{W}_{SI}) + \eta\mathcal{R}_C(\mathbf{W}_I) + \lambda\mathcal{R}_A(\mathbf{H}_I) \quad (8)$$

where $\mathcal{D}_{\mathrm{KL}}(\cdot)$ is the KL-divergence given by (1) and $\eta, \lambda > 0$ are regularization coefficients. The regularization terms are given as:

$$\mathcal{R}_P(\mathbf{W}_{SI}) = \sum_k \sum_m \left[(\alpha_{km} - 1) \ln(w^{SI}_{km}) - \beta_{km} w^{SI}_{km} \right] \quad (9)$$

$$\mathcal{R}_C(\mathbf{W}_I) = \| [\mathbf{W}_S \ \mathbf{W}_N]^T \mathbf{W}_{NI} \|_1 \quad (10)$$

$$\mathcal{R}_A(\mathbf{H}_I) = \| \mathbf{H}_I \|_1 \quad (11)$$

where $\| \cdot \|_1$ denotes the l_1-norm. The term $\mathcal{R}_P(\mathbf{W}_{SI})$ is the logarithm of the prior of $\mathbf{W}_{SI}$ based on (7), where we assume that each entry w^{SI}_{km} is drawn independently and discard irrelevant terms which do not depend on w^{SI}_{km} [16]. The regularization term $\mathcal{R}_C(\mathbf{W}_I)$ accounts for the orthogonality between $[\mathbf{W}_S \ \mathbf{W}_N]$ and $\mathbf{W}_{NI}$. The term $\mathcal{R}_A(\mathbf{H}_I)$ is added for sparse regularization, which implies that a restricted basis vectors will represent the magnitude spectrum dominantly and hence, known to be effective to train the so called parts-based features [20]. The update rules are obtained as:

$$w^I_{km} \leftarrow \begin{cases} \dfrac{(a_{km} - 1) + \sum_q c^I_{kmq}}{\beta_{km} + \sum_q h^I_{mq}}, & 1 \le m \le M_{SF} \\ \dfrac{\sum_q c^I_{kmq}}{\sum_q h^I_{mq} + \eta(\sum_{m'} w^S_{km'} + \sum_{m'} w^N_{km'})}, & \text{else} \end{cases} \quad (12)$$

$$h^I_{mq} \leftarrow \frac{\sum_k c^I_{kmq}}{\sum_k w^I_{km} + \lambda} \quad (13)$$

where c^I_{kmq} is defined as:

$$c^I_{kmq} = \frac{v^I_{kq} w^I_{km} h^I_{mq}}{\sum_{m'} w^I_{km'} h^I_{m'q}}. \quad (14)$$

These update rules are derived by using the majorization-minimization (MM) algorithm which is an iterative optimization method exploiting the convexity of a function to find the maxima or minima [5]. The MM method can be considered as a generalized version of the expectation-maximization (EM) algorithm and hence, guarantees convergence. As a brief interpretation, computing c^I_{kmq} given in (14) corresponds to the expectation-step, whereas the update rules given in (12) and (13) correspond to the maximization-step in the EM algorithm.

At the end of each iteration, a normalization step is included as introduced in Section 2[1]. As for the initialization, we use the free basis matrices obtained at the previous frame, i.e., $\mathbf{W}^F_{l-1} = [\mathbf{W}^{SF}_{l-1} \ \mathbf{W}^{NF}_{l-1}]$, for $\mathbf{W}_I$ and generate positive random numbers for $\mathbf{H}_I$.

2) Basis correction: The estimated $\mathbf{W}_{SI}$ and $\mathbf{W}_{NI}$ may respectively contain some characteristics of the noise and clean speech and therefore, it is necessary to further correct these instantaneous free basis matrices. As a possible approach, the authors in [14] propose to update the basis matrices via temporal smoothing using the SPP as a smoothing factor, where the concept can be intuitively interpreted as follows. If the current time frame mostly contains noise, the current clean speech basis matrix should be composed mostly of the clean speech matrix obtained at previous time frame. We adopt this idea in the proposed framework. Specifically, once $\mathbf{W}_{SI}$ and $\mathbf{W}_{NI}$ are estimated, $\mathbf{W}^{SF}_l$ and $\mathbf{W}^{NF}_l$ are obtained as,

$$w^{SF}_{km,l} = (1 - \zeta_{kl}) w^{SF}_{km,l-1} + \zeta_{kl} w^{SI}_{km} \quad (15)$$

$$w^{NF}_{km,l} = \zeta_{kl} w^{NF}_{km,l-1} + (1 - \zeta_{kl}) w^{NI}_{km} \quad (16)$$

[1] In a strict sense, this type of normalization affects the cost function when the regularization terms are added. However, we use such normalization as usually performed in practical implementation.

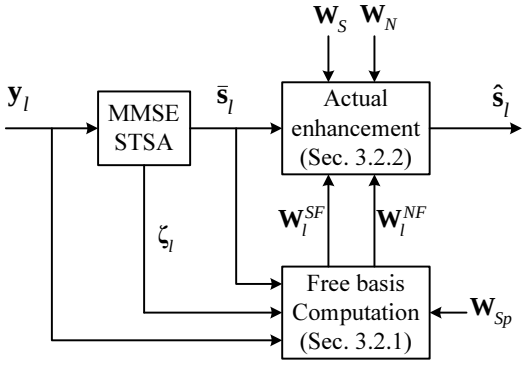

Fig. 1. Simplified block diagram of the proposed N-BC method.

where $\zeta_{kl} \triangleq P(\mathcal{H}_1 | Y_{kl})$ is the posterior SPP computed during the pre-processing stage based on [18] and $\mathcal{H}_1$ indicates the hypothesis of the speech presence.

3.2.2. Actual enhancement

After estimating $\mathbf{W}^{SF}_l$ and $\mathbf{W}^{NF}_l$, we apply the actual enhancement to the pre-processed speech $\bar{\mathbf{s}}_l$. By fixing $\mathbf{W}^Y_l = [\mathbf{W}_S \ \mathbf{W}^{SF}_l \ \mathbf{W}_N \ \mathbf{W}^{NF}_l] \in \mathbb{R}_+^{K \times M_Y}$ where $M_Y = M_S + M_{SF} + M_N + M_{NF}$, the activation vector, $\mathbf{h}^Y_l \in \mathbb{R}_+^{M_Y \times 1}$, is estimated by applying the activation update to $|\bar{\mathbf{s}}_l|$ with sparse regularization. That is,

$$\mathbf{h}^Y_l \leftarrow \mathbf{h}^Y_l \otimes \frac{(\mathbf{W}^Y_l)^T (|\bar{\mathbf{s}}_l| / (\mathbf{W}^Y_l \mathbf{h}^Y_l))}{(\mathbf{W}^Y_l)^T \mathbf{1}_{KI} + \lambda}. \quad (17)$$

The enhanced speech spectrum at the l-th time frame is then estimated using the gain function in (3) along with the parameters given by (4) and the PSDs in (5) and (6). Note that, the PSDs of the clean speech and noise are computed based on $[\mathbf{W}_S \ \mathbf{W}^{SF}_l]$ and $[\mathbf{W}_N \ \mathbf{W}^{NF}_l]$ and their corresponding activation vectors, respectively. A simplified block diagram of the proposed method is illustrated in Figure 1. The proposed NMF algorithm with basis compensation will be referred to as N-BC.

4. EXPERIMENTS

4.1. Methodology

We used clean speech from the TSP database [21] and noise from the NOISEX database [22], where the sampling rate of all signals was adjusted to 16 kHz. The magnitude spectrum of each signal was obtained by using a Hanning window of 512 samples with 75% overlap. For the clean speech, 20 speakers (10 males and 10 females) were considered, whereas the Factory 1, Buccaneer 1 and Hfchannel noises were selected. We considered a speaker-independent application, i.e., one universal basis matrix covering all speakers. The training data for the clean speech was constructed by selecting one sentence from each speaker for a total of 20 sentences (50 seconds), whereas 30 seconds signals were used for the noises. Each of the validation and test speech signals consisted of 6 seconds (2 sentences) signals. All these training, validation and test data were disjoint. The noisy speech was generated from the test and validation signals, respectively, by adding the noise to the clean speech to obtain input SNR of 0, 5 and 10 dB.

For the parameters in the pre-processing, we used a smoothing factor of 0.98 in the decision-directed method for the *a priori* SNR estimation [1], whereas the the smoothing factor of 0.8 was used for the noise PSD estimation [18]. We used $M_S = M_N = 60$

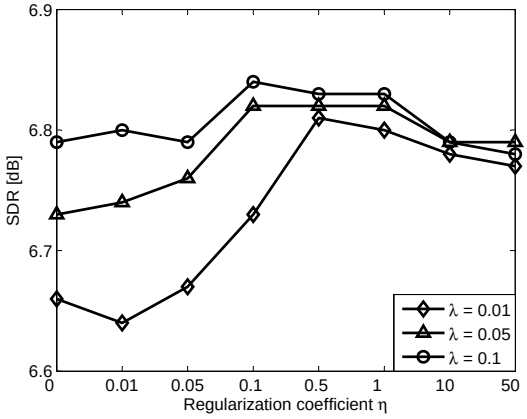

Fig. 2. Average SDR values of the enhanced speech from Factory 1 noise at 0 dB input SNR (validation set).

and $M_{SF} = M_{NF} = 20$ basis vectors, and 20 iterations for each free basis computation and actual enhancement. The smoothing factors for the PSDs estimation in (5) and (6) were selected as $(\tau_S, \tau_N) = (0.4, 0.9)$. The regularization coefficients, λ and η, were determined by observing the performance using the validation set. We considered various choices of $\lambda \in \{0.01, 0.05, 0.1\}$ and $\eta \in \{0, 0.01, 0.05, 0.1, 0.5, 1, 10, 50\}$. Figure 1 shows the average results of the SDR metric for these λ and η values, where the noisy speech was generated by adding the Factory 1 noise to the clean speech at 0 dB input SNR. Based on this observation, as well as similar patterns for the Buccaneer 1 and Hfchannel noises, we ultimately chose $\lambda = 0.1$ and $\eta = 0.1$ for the experiments.

4.2. Results

We used PESQ [23], SDR [24] and SSNR as the objective measures, where a higher value indicates a better results. As for the comparison, we implemented the standard NMF method (NMF) described in Section 2, and two NMF algorithms with basis adaptation (N-BA), where we will refer to each algorithm using its reference number. For these benchmark algorithms, we used the same total number of basis vectors as in the proposed algorithm, i.e., $M_S = M_N = 80$. Although the MMSE log-spectral amplitude estimator (LSA) [25] was used as the pre-processor in [14], we implemented here the MMSE STSA estimator which is used in the proposed N-BC method. Table 1 shows the average results of using the matched noise basis vectors, where we employed the same type of the noise basis vectors as the noise observed in the noisy speech. Table 2 shows the results of using mismatched noise basis vectors, where we evaluated all algorithms with the noise basis vectors obtained from the Babble noise training data (which is also included in the NOISEX database). We can see from Table 1 that the proposed method yielded the best results, in general. Furthermore, we can observe in Table 2 that the proposed algorithm gave even much better improvements for the mismatched case. Specifically, comparing the results between the matched and mismatched cases, we see that performance of the proposed N-BC method decreased much less than the performance of the benchmark algorithms. Similar results were found for 10 dB input SNR. It is thus verified that using the free basis vectors captures the unobserved features in the training data better than updating the complete set of basis vectors.

Informal listening tests were also conducted. In general, the proposed method offered the best enhanced speech quality compared to the others, in terms of the speech distortion and especially, the noise reduction. Among the benchmark algorithms, [14] provided reasonable results since it removed a significant noise components. However, the clean speech was also attenuated significantly. Moreover, for the mismatched case, it failed to properly capture the features corresponding to the ringing sound at high frequencies in the Buccaneer 1 noise. When only considering the proposed method, the enhanced speech was slightly attenuated compared to the clean speech. This could be handled by further applying frequency weights or using a more effective basis matrix estimation algorithm during the training stage, e.g., discriminative training criteria, which will be considered in our future work.

Table 1. Average results with matched noise basis vectors

Input SNR		Eval.	Noisy	NMF	N-BA [13]	N-BA [14]	N-BC
Factory 1	0 dB	PESQ	1.34	1.61	1.71	1.90	**1.92**
		SDR	0.06	3.91	4.30	**6.93**	6.83
		SSNR	-4.34	-1.52	-0.66	0.72	**0.85**
	5 dB	PESQ	1.71	2.01	2.06	2.25	**2.29**
		SDR	5.04	8.87	7.64	9.48	**10.29**
		SSNR	-1.14	1.69	1.53	2.85	**3.38**
Buccaneer 1	0 dB	PESQ	1.20	1.56	1.68	1.91	**2.01**
		SDR	0.05	4.03	4.87	7.55	**7.71**
		SSNR	-4.54	-1.75	-0.57	0.89	**1.14**
	5 dB	PESQ	1.54	1.94	1.99	2.27	**2.36**
		SDR	5.04	8.91	8.06	10.04	**10.71**
		SSNR	-1.37	1.51	1.83	3.20	**3.51**
Hfchannel	0 dB	PESQ	1.17	1.51	1.69	1.91	**2.01**
		SDR	0.06	4.93	6.72	8.72	**8.83**
		SSNR	-4.53	-1.24	0.63	2.22	**2.36**
	5 dB	PESQ	1.44	1.85	2.00	2.28	**2.37**
		SDR	5.04	9.69	9.57	11.13	**11.79**
		SSNR	-1.36	2.03	3.02	4.58	**4.80**

Table 2. Average results with mismatched noise basis vectors

Input SNR		Eval.	Noisy	NMF	N-BA [13]	N-BA [14]	N-BC
Factory 1	0 dB	PESQ	1.34	1.48	1.37	1.78	**1.87**
		SDR	0.06	2.83	2.04	5.94	**6.46**
		SSNR	-4.43	-1.80	-1.29	0.32	**0.74**
	5 dB	PESQ	1.71	1.88	1.70	2.09	**2.25**
		SDR	5.04	7.67	5.21	7.50	**9.69**
		SSNR	-1.14	0.97	0.45	2.02	**2.95**
Buccaneer 1	0 dB	PESQ	1.20	1.32	1.31	1.60	**1.84**
		SDR	0.05	0.97	1.49	3.96	**7.20**
		SSNR	-4.54	-3.09	-1.71	-1.21	**0.85**
	5 dB	PESQ	1.54	1.67	1.63	1.89	**2.23**
		SDR	5.04	5.96	4.77	6.58	**10.07**
		SSNR	-1.37	-0.25	0.07	0.83	**3.02**
Hfchannel	0 dB	PESQ	1.17	1.25	1.27	1.46	**1.86**
		SDR	0.06	-0.06	0.56	5.19	**8.15**
		SSNR	-4.53	-3.59	-2.32	0.33	**1.94**
	5 dB	PESQ	1.44	1.52	1.49	1.87	**2.24**
		SDR	5.04	5.20	4.43	7.21	**10.97**
		SSNR	-1.36	-0.68	-0.29	2.30	**4.12**

5. CONCLUSION

A basis compensation algorithm of the NMF model for supervised speech enhancement has been proposed. We used free basis vectors for both the clean speech and noise during the enhancement stage in order to capture the features which are not included in the training data. The free basis vectors were estimated by exploiting *a prior* knowledge and orthogonality regularization for the clean speech and noise, respectively. Experiments showed that the proposed method provided better results than the benchmark algorithms.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126-137, Mar. 1999.

[3] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 700-708, Nov. 2003.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, pp. 556-562, 2001.

[5] C. Fevotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421-2456, Mar. 2011.

[6] G. Buchsbaum and O. Bloch, "Color categories revealed by non-negative matrix factorization of munsell color spectra," *Vision Research*, vol. 42, no. 5, pp. 559-563, Mar. 2002.

[7] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness constraint," *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, no. 3, pp. 1066-1074, Mar. 2007.

[8] N. Mohammadiha, P. Smaragdis and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech and Language Process.*, vol. 21, no. 10, pp. 2140-2151, Oct. 2013.

[9] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. ICASSP*, pp. 17-20, May 2011.

[10] H. Chung, E. Plourde and B. Champagne, "Regularized NMF-based speech enhancement with spectral componenets modeled by Gaussian mixtures," in *Proc. MLSP*, pp. 1-6, Sep. 2014.

[11] T. Virtanen and A. T. Cemgil, "Mixtures of Gamma priors for non-negative matrix factorization based speech separation," in *Proc. Independent Component Analysis and Signal Separation*, pp. 646-653, Mar. 2009.

[12] M. A. Carlin, N. Malyska and T. F. Quatieri, "Speech enhancement using sparse convolutive non-negative matrix factorization with basis adaptation," in *Proc. Interspeech*, pp. 583-586, Sep. 2011.

[13] E. M. Grais and H. Erdogan, "Adaptation of speaker-specific bases in non-negative matrix factorization for single channel speech-music separation," in *Proc. Interspeech*, Aug. 2011.

[14] K. Kwon, J. W. Shin and N. S. Kim, "NMF-based speech enhancement using bases update," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 450-454, Apr. 2015.

[15] K. Yagi, Y. Takahashi, H. Saruwatari, K. Shikano and K. Kondo, "Music signal separation by orthogonality and maximum-distance constrained nonnegative matrix factorization with target signal information," in *Proc. Audio Eng. Society 45th International Conference*, pp. 2-5, Mar. 2012.

[16] T. Virtanen, A. T. Cemgil and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. ICASSP*, pp. 1825-1828, Mar. 2008.

[17] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," in *Proc. Interspeech*, Aug. 2013.

[18] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383-1393, May 2012.

[19] 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," in *Proc. Independent Component Analaysis and Signal Separation*, pp. 338-346, Mar. 2012.

[20] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraint," *Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, Nov. 2004.

[21] P. Kabal, *TSP Speech Database*. Tech. Rep., McGill University, Montreal, Canada, 2002.

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

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

[24] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, no. 4, pp. 1462-1469, July 2006.

[25] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-33, no. 2, pp. 443-445, Apr. 1985.