SINGLE-CHANNEL ENHANCEMENT OF CONVOLUTIVE NOISY SPEECH BASED ON A DISCRIMINATIVE NMF ALGORITHM

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 introduce a discriminative training algorithm of the non-negative matrix factorization (NMF) model for single-channel enhancement of *convolutive* noisy speech. The basis vectors for the clean speech and noises are estimated simultaneously during the training stage by incorporating the concept of classification from machine learning. Specifically, we employ the probabilistic generative model (PGM) of classification, specified by an inverse Gaussian distribution, as *a priori* structure for the basis vectors. Both the NMF and classification parameters are obtained by using the expectation-maximization (EM) algorithm, which guarantees convergence to a stationary point. Experimental results show that the proposed algorithm provides better enhancement performance than the benchmark algorithms.

Index Terms— Single-channel speech enhancement, non-negative matrix factorization, discriminative training, probabilistic generative model, classification

1. INTRODUCTION

Numerous algorithms for single-channel speech enhancement, aiming at removing the background noise from a noisy speech, have been proposed in the past: such as spectral subtraction [1], minimum mean-square error (MMSE) estimator [2] or subspace decomposition [3]. These classical methods, however, tend to provide limited performance in adverse noisy environments, e.g., low input signal-to-noise ratio (SNR) or non-stationary noise conditions. Recently, non-negative matrix factorization (NMF) methods have been successfully applied to diverse problems including source separation [4] and speech enhancement [5]. In general, NMF is a dimensionality reduction tool that decomposes a given matrix into basis and activation matrices with non-negative elements constraint [6]. In audio and speech 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.

Most existing single-channel source separation or speech enhancement algorithms consider an instantaneous mixture, i.e., the noisy speech is obtained by simply adding the anechoic background noise to the clean speech. An extension of the conventional NMF model, known as convolutive NMF (CNMF) [7] has been proposed to effectively capture the time-varying characteristics of the audio or speech signals, and has been applied to speech separation [7] and speech enhancement [8, 9] problems. However, the term convolutive in CNMF indicates that the given spectrum is modeled as a shifted sum of time-varying basis matrix and hence, these algorithms do not consider the explicit convolutive mixing process specified by a mixing filter such as room impulse response (RIR).

Another problem of the NMF-based framework is that the basis vectors of the different sources may share similar characteristics. For example, the basis vectors of speech spectrum can represent the noise spectrum and hence, the enhanced speech may contain noise components that have similar features to the clean speech. Recently, several discriminative training algorithms of the NMF model with application to source separation or speech enhancement for instantaneous mixtures have been proposed to solve this problem, in which the goal is to train the basis vectors of each source in a way that prevents them from representing each other (see [10] and references therein). However, such training criteria have not been yet employed for convolutive noisy speech.

In this paper, we introduce a discriminative training algorithm of NMF model for single-channel enhancement of convolutive noisy speech, which is an extension of our previous work in [10], where the main idea was to estimate the basis matrices during the training stage by constraining them to belong to one of several classes. To this end, we considered a traditional Gaussian-distributed probabilistic generative model (PGM) of classification [11] along with the NMF model [12, 13]. In this paper, we explicitly formulate and exploit the convolutive signal model motivated by [12], and instead employ an inverse Gaussian distribution as the PGM for classification to bring more coherence into the NMF model. The update rules of the NMF model and the PGM parameters for classification are jointly estimated via the expectation-maximization (EM) algorithm. Experimental results show that the proposed algorithm provides better enhancement performance than the benchmark algorithms.

2. SIGNAL MODEL

The convolutive noisy speech can be expressed in the short-time Fourier transform (STFT) domain as [12]

$$Y_{kl} = \mathbf{A}_k \mathbf{x}_{kl} + B_{kl} \tag{1}$$

where Y_{kl} is the complex-valued STFT of the convolutive noisy speech, $\mathbf{x}_{kl} = [S_{kl}\, N_{kl}]^T \in \mathbb{C}^{2\times 1}$ is a point source vector consisting of the clean speech and noise, S_{kl} and N_{kl}, $\mathbf{A}_k = [A_k^S\ A_k^N] \in \mathbb{C}^{1\times 2}$ is a vector of the mixing filters (e.g., RIRs which model the paths from the clean speech and noise to the microphone), B_{kl} is a residual error (independent of $\mathbf{x}_{kl}$), and $k = \{1, ..., K\}$ and $l = \{1, ..., L\}$ are the frequency and time frame indices. The residual error, which is shown to prevent the EM algorithm from potential numerical instabilities and slow convergence [12], can be modeled by a stationary Gaussian random process with zero-mean and variance σ_k^2 [12, 14].

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

The goal is to recover the clean speech in either the form of the point source S_{kl} or the so-called *image source* $Z^S_{kl} = A^S_k S_{kl}$ [12, 14]. In this paper, we consider the latter case to evaluate the enhancement performance. In the following, we introduce two underlying PGMs which will be employed in the proposed framework: NMF and classification models.

2.1. NMF model

For a given matrix $\mathbf{V} \in \mathbb{R}_+^{K\times L}$, NMF finds a local optimal decomposition of $\mathbf{V} \approx \mathbf{WH}$, 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, $\mathbb{R}_+$ denotes the set of non-negative real numbers, and M is the number of basis vectors. The factorization is obtained by minimizing a suitable cost function, such as Kullback-Leibler (KL) [6] or Itakura-Saito (IS) [13] divergence. In this paper, we consider the IS-divergence since it is known to provide a desirable statistical interpretation of the audio and speech signals [12, 13]. Moreover, we can explicitly employ the complex-valued spectrum which is necessary to handle the convolutive signal model given by (1).

Within a statistical framework, the complex-valued observation X_{kl} is assumed to be a sum of M latent variables, c^m_{kl}, as

$$X_{kl} = \sum_{m=1}^{M} c^m_{kl}, \qquad c^m_{kl} \sim \mathcal{N}_c(0, w_{km}h_{ml}) \tag{2}$$

where $\mathcal{N}_c(\mu, \sigma^2)$ is a complex Gaussian distribution with mean μ and variance σ^2. Assuming that the latent variables are mutually independent, it has been shown that maximizing the log-likelihood function (LLF) based on the model (2) with respect to w_{km} and h_{ml} is equivalent to minimizing the IS divergence [13].

2.2. Classification model

In the classification problem, the input vector $\mathbf{w} = [w_k] \in \mathbb{R}^K$ under test is assigned to one of I classes. The goal is to find a partition of the observation space into decision regions that will minimize the classification error, by using training data and their corresponding class labels. Among various approaches to solve the classification problem (e.g, PGM and discriminative modeling [11]), we consider the PGM since it can provide the necessary *a priori* distributions to be used in the proposed framework[1].

By ignoring possible correlations between different entries in $\mathbf{w}$, the class-conditional density based on the inverse-Gaussian distribution can be expressed as $p(\mathbf{w}|d_i = 1) = \prod_{k=1}^K \mathcal{IN}(\mu^i_k, \lambda_k)$ where $d_i \in \{0, 1\}$ is a target class label for the class $i \in \{0, ..., I-1\}$ and

$$\mathcal{IN}(\mu, \lambda) = \left(\frac{\lambda}{2\pi w^3}\right)^{1/2} \exp\left[\frac{-\lambda(w-\mu)^2}{2\mu^2 w}\right] \tag{3}$$

is the inverse-Gaussian distribution defined for a positive value ($w > 0$) with mean μ and shape parameter λ.

Suppose we have a training set $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_M]$ and $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_M]$, where $\mathbf{d}_m = [d_{im}]$ with $d_{im} \in \{0, 1\}$ is an $I \times 1$ target class label vector such that $\sum_i d_{im} = 1$. Assuming that the columns $\mathbf{w}_m$ are independently drawn, the likelihood function is given by

$$p(\mathbf{W}, \mathbf{D}; \boldsymbol{\theta}_C) = \prod_{m=1}^{M}\prod_{i=0}^{I-1} [p(\mathbf{w}_m|d_i = 1)p_i]^{d_{im}} \tag{4}$$

[1]The term *discriminative training* used in this paper differs from *training discriminative model*, where the latter aims at maximizing the posterior distribution. Although some authors consider both terms equivalently (e.g., [15, 16]), we refer to the former as a training method aiming at estimating arbitrary parameters to be distinct (e.g., [17, 18]).

where $\boldsymbol{\theta}_C = \{\{p_i, \{\mu^i_k\}\}_{i=0}^{I-1}, \{\lambda_k\}\}$ is a PGM parameter set for classification and $p_i \triangleq p(d_i = 1)$ is the prior class probability. The set $\boldsymbol{\theta}_C$ can be estimated via the maximum likelihood (ML) criterion.

3. PROPOSED ALGORITHM

In this section, we first explicitly address the prior structures for the PGM in (1), which will be used in the proposed framework. Subsequently, we explain the proposed training and enhancement stages.

3.1. Prior structures

We denote by M_i the number basis vectors in class i (such that $M = \sum_i M_i$), and by L_i the number of time frames in class i. For the basis vectors, the log-likelihood in (4) can be simply rearranged as

$$p(\mathbf{W}; \boldsymbol{\theta}_C) = \prod_{i=0}^{I-1}\prod_{m=1}^{M_i} \left[p(\mathbf{w}^i_m|d_i = 1)p_i\right]. \tag{5}$$

where $p(\mathbf{w}^i_m|d_i = 1)$ is given in (4), and we omit the dependence on $\mathbf{D}$ in $p(\mathbf{W}, \mathbf{D}; \boldsymbol{\theta}_C)$ hereafter for convenience.

For the activations, we employ sparse NMF regularization, which can be implemented by modeling the entries of $\mathbf{H}$ with an exponential distribution within a statistical framework [19]. Assuming that the entries are independent and identically distributed, the prior of $\mathbf{H}$ can be written as

$$p(\mathbf{H}; \eta) = \prod_{i=0}^{I-1} \eta^{M_i L_i} \exp\left[-\eta \sum_{m=1}^{M_i}\sum_{l=1}^{L_i} h^i_{ml}\right] \tag{6}$$

where the parameter η controls the degree of sparsity.

3.2. Training stage

In the proposed framework, we use the class index $i = 0$ for the clean speech and $i = 1, ..., I-1$ for the different noise types. Let us denote by $\mathbf{Z}^i = [Z^i_{kl}]$ and $\mathbf{X}^i = [X^i_{kl}]$ the i-th image and point source spectra, respectively. For given training data sets of the clean speech and noise image spectra $\mathbf{Z} = \{\mathbf{Z}^i\}$, our goal is to estimate $\boldsymbol{\theta} = \{\{A^i_k\}, \{w^i_{km}\}, \{h^i_{ml}\}\}$ and $\boldsymbol{\theta}_C$ jointly. The complete-data LLF can be expressed as

$$\begin{aligned}
&\ln p(\mathbf{Z}, \mathbf{C}, \mathbf{W}, \mathbf{H}; \boldsymbol{\theta}_C, \eta) \\
&\quad = \ln p(\mathbf{Z}|\mathbf{C}) + \ln p(\mathbf{C}|\mathbf{W}, \mathbf{H}) + \ln p(\mathbf{W}; \boldsymbol{\theta}_C) + \ln p(\mathbf{H}; \eta) \\
&\quad \stackrel{c}{=} -\sum_{i=0}^{I-1}\sum_{k=1}^{K}\sum_{l=1}^{L_i} \left[\ln(\sigma^i_k)^2 + |Z^i_{kl} - A^i_k X^i_{kl}|^2(\sigma^i_k)^{-2}\right] \\
&\qquad - \sum_{i=0}^{I-1}\sum_{k=1}^{K}\sum_{l=1}^{L_i}\sum_{m=1}^{M_i} \left[\ln(w^i_{km}h^i_{ml}) + \frac{|c^{m,i}_{kl}|^2}{w^i_{km}h^i_{ml}}\right] \\
&\qquad + \sum_{i=0}^{I-1}\sum_{k=1}^{K}\sum_{m=1}^{M_i} \left[\frac{1}{2}(\ln\lambda_k - 3\ln w^i_{km}) - \frac{\lambda_k(w^i_{km} - \mu^i_k)^2}{2(\mu^i_k)^2 w^i_{km}}\right] \\
&\qquad + \sum_{i=0}^{I-1} \left[M_i L_i \ln\eta - \eta \sum_{m=1}^{M_i}\sum_{l=1}^{L_i} h^i_{ml}\right]
\end{aligned} \tag{7}$$

where $\mathbf{C} = \{c^{m,i}_{kl}\}$ is the set of latent variables defined in (2), and $\stackrel{c}{=}$ indicates equality up to a constant term.

Application of the EM algorithm to (7) consists of two stages: i) *expectation step (E-step)*, computing the posterior distribution of the latent variable given the observation and the expectation of the sufficient statistics accordingly, and ii) *maximization step (M-step)*, estimating the parameters by maximizing the conditional expectation of

the complete-data LLF with respect to the posterior distribution (i.e., $\mathcal{L}_C(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int \ln p(\mathbf{Z}, \mathbf{C}|\boldsymbol{\theta}) p(\mathbf{C}|\mathbf{Z}, \boldsymbol{\theta}') d\mathbf{C}$, where $\boldsymbol{\theta}'$ is the parameter estimated in the previous EM iteration). Defining $u_{kl}^{m,i} \triangleq |c_{kl}^{m,i}|^2$, the E-step is found to be [12]:

$$\hat{u}_{kl}^{m,i} = |\hat{c}_{kl}^{m,i}|^2 + (1 - G_{c,kl}^i) w_{km}^i h_{ml}^i \tag{8}$$

$$\hat{R}_{zx,k}^i = \frac{1}{L_i} \sum_{l=1}^{L_i} Z_{kl}^i (\hat{X}_{kl}^i)^* \tag{9}$$

$$\hat{R}_{xx,k}^i = \frac{1}{L_i} \sum_{l=1}^{L_i} \left[|\hat{X}_{kl}^i|^2 + (1 - G_{x,kl}^i A_k^i) v_{kl}^i \right] \tag{10}$$

where * refers to a conjugate operation, $v_{kl}^i = \sum_{m=1}^{M_i} w_{km}^i h_{ml}^i$ and

$$\hat{X}_{kl}^i = G_{x,kl}^i Z_{kl}^i, \qquad G_{x,kl}^i = v_{kl}^i (A_k^i)^* (\Sigma_{z,kl}^i)^{-1} \tag{11}$$

$$\hat{c}_{kl}^{m,i} = G_{c,kl}^i Z_{kl}^i, \qquad G_{c,kl}^i = (w_{kl}^i h_{ml}^i)(A_k^i)^* (\Sigma_{z,kl}^i)^{-1} \tag{12}$$

$$\Sigma_{z,kl}^i = |A_k^i|^2 v_{kl}^i + (\sigma_k^i)^2. \tag{13}$$

The M-step is as follows. The mixing filter is found to be

$$A_k^i = \hat{R}_{zx,k} \hat{R}_{xx,k}^{-1}. \tag{14}$$

The basis elements are found by setting the partial derivative of $\mathcal{L}_C(\boldsymbol{\theta}|\boldsymbol{\theta}')$ with respect to w_{km}^i to zero, which leads to solving the following second-order polynomial equation:

$$(w_{km}^i)^2 + \underbrace{\frac{(2+3L_i)(\mu_k^i)^2}{\lambda_k}}_{\triangleq\, q_{w1}^i} w_{km}^i - \underbrace{(\mu_k^i)^2 \left(1 + \frac{2}{\lambda_k} \sum_{l=1}^{L_i} \frac{\hat{u}_{kl}^{m,i}}{h_{ml}^i} \right)}_{\triangleq\, q_{w2}^i} = 0. \tag{15}$$

Hence, the resulting update rule of w_{km}^i is found to be

$$w_{km}^i = \frac{2 q_{w2}^i}{q_{w1}^2 + \sqrt{(q_{w1}^i)^2 + 4 q_{w2}^i}} \tag{16}$$

Following a similar approach as for the basis estimation, the update rule of h_{ml}^i is obtained as

$$h_{ml}^i = \frac{2 q_{h2}^i}{K + \sqrt{K^2 + 4\eta q_{h2}^i}} \tag{17}$$

where $q_{h2}^i \triangleq \sum_{k=1}^{K} (\hat{u}_{kl}^{m,i} / w_{km}^i)$. The residual noise variance, $(\sigma_k^i)^2$, also can be estimated by maximizing $\mathcal{L}_C(\boldsymbol{\theta}|\boldsymbol{\theta}')$. However, we instead follow a strategy called *simulated annealing with noise injection* method introduced in [12], since it is shown to provide faster convergence of the EM iteration. Specifically, the residual noise variance is initialized with an average channel empirical variance divided by 100 (i.e., $(\sigma_k^i)^2 = \sum_l |Z_{kl}^i|^2 / (100 L_i)$), and is gradually decreased through iterations to a small value, e.g., 1e-10. A random noise is added to $\mathbf{Z}^i$ at each EM iteration, accordingly.

The hyper-parameter set $\boldsymbol{\theta}_C$ is estimated by maximizing the marginal likelihood $p(\mathbf{Z}|\mathbf{H}; \boldsymbol{\theta}_C) = \int p(\mathbf{Z}, \mathbf{W}|\mathbf{H}; \boldsymbol{\theta}_C) d\mathbf{W}$. Assuming that $\mathbf{W}$ is *well-determined*, maximizing the marginal likelihood becomes equivalent to maximizing (7) [10, 11]. Consequently, the set $\boldsymbol{\theta}_C$ is simply found by applying the ML criterion to (5), where the resulting estimate in a closed form is interleaved with the EM update, as

$$\hat{\mu}_k^i = \frac{1}{M_i} \sum_{m=1}^{M_i} w_{km}^i, \quad \frac{1}{\hat{\lambda}_k} = \frac{1}{M} \sum_{i=0}^{I-1} \sum_{m=1}^{M_i} \left(\frac{1}{w_{km}^i} - \frac{1}{\hat{\mu}_k^i} \right) \tag{18}$$

and $p_i = M_i / M$.

To prevent scale indeterminacies, we add a normalization step by adopting the strategies in [12] and [20]. That is, after computing (14) and (16), we normalize A_k^i by its magnitude $|A_k^i|$ and scale w_{km}^i accordingly, and then compute (17). As for initialization, we generate random complex numbers for A_k^i. For the basis and activations, we apply the standard multiplicative update (MU) rules based on KL-divergence [6] to the magnitude-square of the image source spectra as in [12] for 10 iterations.

3.3. Enhancement stage

During the enhancement stage, by concatenating and fixing the basis matrices of the clean speech and noise obtained during the training stage as $\mathbf{W} = [\mathbf{W}_S \ \mathbf{W}_N]$, we estimate the mixing filters, $\mathbf{A}_k$, and activation matrix, $\mathbf{H} = [\mathbf{H}_S^T \ \mathbf{H}_N^T]^T$, from the convolutive noisy speech $\mathbf{Y}$. The parameter estimation via the EM algorithm can be derived similarly as in the training stage. Based on the signal model in (1), the necessary sufficient statistics corresponding to (9)-(10) and the mixing filter in (14) take either a vector or matrix form. A detailed expression for the paramter estimation can be found in [12], where the activation matrix $\mathbf{H}$ is estimated by (17) in the proposed framework. Once the parameters are obtained, we estimate the image spectrum of the clean speech using (11), (13) and $Z_{kl}^S = A_k^S S_{kl}$, where we ignore the small value of the residual noise variance σ_k^2. Moreover, since the mixing filter is normalized, the estimated image spectrum of the clean speech can be written as

$$\hat{Z}_{kl}^S = \frac{\hat{p}_{kl}^S}{\hat{p}_{kl}^S + \hat{p}_{kl}^N} Y_{kl} \tag{19}$$

where $\hat{p}_{kl}^S$ and $\hat{p}_{kl}^S$ 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 [10, 21]

$$\hat{p}_{kl}^S = \tau_S \hat{p}_{k,l-1}^S + (1 - \tau_S) \sum_{m=1}^{M_S} w_{km}^S h_{ml}^S \tag{20}$$

$$\hat{p}_{kl}^N = \tau_N \hat{p}_{k,l-1}^N + (1 - \tau_N) \sum_{m=1}^{M_N} w_{km}^N h_{ml}^N \tag{21}$$

where τ_S and τ_N are the smoothing factors for the clean speech and noise, respectively. Finally, the enhanced image speech signal in the time-domain is reconstructed by applying the inverse STFT followed by the overlap-add method. Note that the set $\boldsymbol{\theta}_C$ can be used for the noise classification using a Bayes' rule in advance to the enhancement [10]. In this case, the additional noise basis vector $\mathbf{w}$ needed for the classification can be obtained through $[\mathbf{W}_S \ \mathbf{w}]$ by applying the standard MU rule to $\mathbf{V} = [|Y_{kl}|^2]$. In this paper, however, we simply assume that the noise type is known *a priori*.

4. EXPERIMENTS

We conducted experiments by considering a rectangular room with dimensions of $4 \times 5 \times 3$ m ($x \times y \times z$) as illustrated in Fig. 1. A microphone and three point sources (P_1, P_2 and P_3) were placed at the elevation of $z = 1.3$ m. The RIRs with respect to different source positions were obtained by using the simulator in [22], where we considered the reverberation time of $T_{60} = 50$ and 200 ms. We used clean speech from the TSP database [23] and noise from the NOISEX database [24], where the sampling rate of all signals was set to 16 kHz. For the clean speech ($i = 0$), 20 speakers (10 males and 10 females) were chosen, whereas the Factory 1 ($i = 1$), Buccaneer

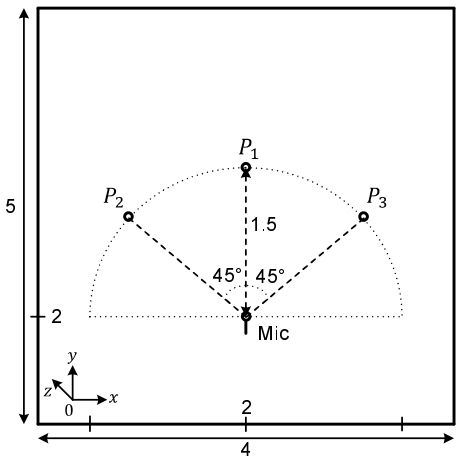

Fig. 1. Room geometry (length in meters, angle in degrees).

1 ($i = 2$), HF-Channel ($i = 3$) and Destroyerops ($i = 4$) were selected[2] The corresponding speech and noise files were divided into two disjoint groups: i) *training data*, used for estimating the basis matrix for each class i during the training stage, and ii) *test data*, used during the enhancement stage to evaluate the enhancement performance. We considered a speaker-independent (SI) application, where one *universal* basis matrix covering all speakers is estimated. To this end, we constructed the training data of the clean speech by selecting 3 sentences per speaker and concatenating them, for a total of 60 sentences (2 minutes long signal), whereas a 2 minutes long signal was used for each type of noise. All training data were located at P_1 and then convolved with the corresponding RIR to obtain the image source signals.

The convolutive noisy speech signals were generated from the test data by summing the image signals of the clean speech and noise. Specifically, we selected three sentences (8 seconds long signal) per speaker for the clean speech, whereas we selected 10 seconds long segments for each noise type. The point sources of the clean speech and noise were located at P_3 and P_2, respectively (see Fig. 1). The image signals of the clean speech and noise were obtained by convloving the point source signals with their corresponding RIRs. Subsequently, the image noise signal was added to the image speech signal to have input SNR of 0 and 5 dB. Regarding the implementation, the STFT of each signal was obtained by using a Hanning window of 512 samples with 75% overlap. We used $M_i = 60$ basis vectors for all i. Sparsity and temporal smoothing factors were selected as $\eta = 5$ and $(\tau_S, \tau_N) = (0.4, 0.9)$.

We considered the perceptual evaluation of speech quality (PESQ) [25] and signal-to-distortion ratio (SDR) [26] as the objective measures, where a higher value indicates a better result. To compare the proposed method, we implemented the standard NMF method in [6] (see [10] for its application to supervised speech enhancement), CNMF method [7] where we used the maximum shift length of 3 for the convolution process of the basis and activation matrices. In addition, we implemented the discriminative training algorithm of the NMF model based on class probabilities in [10], which will be referred to as DCP. Basic settings such as the STFT analysis and synthesis process, the number of basis vectors and temporal smoothing factors were kept identical for fair comparison. The average results over all speakers for $T_{60} = 50$ and 200 ms are respectively shown in Table 1 and 2. We can see that the proposed method provided better results than the benchmark algorithms under considered input SNRs, except in specific case, e.g., SDR value for the Factory 1 noise at 0 dB input SNR. It also can be seen that the performance for $T_{60} = 200$ ms has been degraded compared to the 50ms. The main reason is that the signal model in (1) is appropriate when the RIR is much shorter than the STFT analysis window length [12, 14]. In order to improve the enhancement performance even for a highly reverberant environment, it would be necessary to consider an extended signal model, e.g., the latent variable c_{kl}^m in (2) modeled by an auto-regressive process [27], which will be considered in our future work.

[2]Although the considered noise types are more likely to originate outside the room, we assume that they are generated from a point source inside the room for a practical simulation.

Table 1. Average results for T_{60} =50 ms

Input SNR		Eval.	Noisy	NMF [6]	CNMF [7]	DCP [10]	Prop.
Fact. 1	0 dB	PESQ	1.40	1.66	1.70	1.75	**1.85**
		SDR	0.04	3.95	4.50	**6.21**	5.95
	5 dB	PESQ	1.76	2.07	2.11	2.10	**2.29**
		SDR	5.03	8.84	9.34	9.96	**10.23**
Bucc. 1	0 dB	PESQ	1.25	1.69	1.73	1.88	**2.06**
		SDR	0.03	4.39	5.10	7.17	**7.53**
	5 dB	PESQ	1.59	2.07	2.11	2.11	**2.43**
		SDR	5.02	9.23	9.68	10.53	**10.98**
HF-Chan.	0 dB	PESQ	1.20	1.69	1.69	1.95	**2.06**
		SDR	0.04	5.38	6.15	8.27	**8.82**
	5 dB	PESQ	1.48	2.04	2.07	2.10	**2.39**
		SDR	5.02	9.95	10.57	11.39	**11.95**
Dest.ops	0 dB	PESQ	1.59	1.89	2.00	1.95	**2.14**
		SDR	0.03	4.15	6.34	6.43	**7.19**
	5 dB	PESQ	1.99	2.29	2.39	2.29	**2.50**
		SDR	5.02	9.04	10.58	9.64	**11.07**

Table 2. Average results for $T_{60} = 200$ ms

Input SNR		Eval.	Noisy	NMF [6]	CNMF [7]	DCP [10]	Prop.
Fact. 1	0 dB	PESQ	1.42	1.67	1.66	1.74	**1.81**
		SDR	0.06	3.32	3.18	**4.97**	4.63
	5 dB	PESQ	1.81	2.08	2.07	2.12	**2.23**
		SDR	5.04	8.12	8.13	9.05	**9.27**
Bucc. 1	0 dB	PESQ	1.35	1.73	1.75	1.89	**2.01**
		SDR	0.04	4.54	5.00	6.79	**7.29**
	5 dB	PESQ	1.71	2.12	2.14	2.22	**2.40**
		SDR	5.03	9.20	9.78	10.27	**11.04**
HF-Chan.	0 dB	PESQ	1.22	1.68	1.59	1.84	**1.93**
		SDR	0.04	5.20	5.12	6.71	**7.71**
	5 dB	PESQ	1.50	2.03	1.95	2.15	**2.26**
		SDR	5.03	9.72	9.85	9.76	**11.20**
Dest.ops	0 dB	PESQ	1.65	1.92	1.95	1.98	**2.05**
		SDR	0.05	4.00	4.18	5.47	**5.51**
	5 dB	PESQ	2.07	2.32	2.34	2.35	**2.44**
		SDR	5.03	8.67	8.96	9.05	**9.84**

5. CONCLUSION

We introduced a discriminative training algorithm of NMF model for single-channel enhancement of convolutive noisy speech. The convolutive signal model has been explicitly formulated and employed in the proposed framework. Moreover, the basis vectors for the clean speech and noises were estimated simultaneously during the training stage by employing the PGM of classification, specified by an inverse Gaussian distribution, as *a priori* structure. Both the NMF and classification parameters were obtained via the EM algorithm. Experimental results under different reverberant conditions showed that the proposed algorithm provides better enhancement performance than the benchmark algorithms.

6. REFERENCES

[1] 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.

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

[3] F. Jabloun and B. Champagne, "Auditory-based spectral amplitude estimator for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 700-708, Nov. 2003.

[4] 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.

[5] 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.

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

[7] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 1-12, Jan. 2007.

[8] 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. 2012.

[9] Z. Chen, B. McFee and D. P. Ellis, "Speech enhancement by low-rank and convolutive dictionary spectrogram decomposition," in *Proc. Interspeech*, pp. 2833-2837, Sep. 2014.

[10] H. Chung, E. Plourde and B. Champagne, "Discriminative training of NMF model based on class probabilities for speech enhancement," *IEEE Signal Process. Letters*, vol. 23, no.4, pp. 502-506, Feb. 2016.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 3, pp. 550-563, Mar. 2010.

[13] C. Févotte, N. Bertin and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793-830, Mar. 2009.

[14] S. Arberet, A. Ozerov, N.Q.K. Duong, E. Vincent, R. Gribonval, F. Bimbot and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. Int. Conf. Information Science, Signal Process. and their Applications*, pp. 1-4, May 2010.

[15] A. Holub and P. Perona, "A discriminative framework for modeling object classes," in Proc. *CVPR*, pp. 664-671, June 2005.

[16] H. Jiang, "Discriminative training of HMMs for automatic speech recognition: A Survey," *Computer Speech and Language*, vol. 24, pp. 589-608, Aug. 2009.

[17] N. Guan, D. Tao, Z. Luo and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2030-2048, July 2011.

[18] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," in Proc. *Interspeech*, pp. 808-812, Aug. 2013.

[19] M. N. Schmidt and J. Larsen, "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *Proc. Workshop on Machine Learning for Signal Process.*, pp. 486-491, Oct. 2008.

[20] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proc. Int. Joint Conf. Neural Networks*, pp. 2529-2533, July 2004.

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

[22] E. A. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. America*, vol. 124, no. 1, pp. 269-277, July 2008.

[23] P. Kabal, *TSP Speech Database*. Tech. Rep., McGill University, Montreal, Canada, 09-02, 2002.

[24] A. Varga and H. J. 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, July 1993.

[25] 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.

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

[27] R. Badeau and M. D. Plumbley, "Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 11, pp. 1670-1680, Nov. 2014.