Discriminative Training of NMF Model Based on Class Probabilities for Speech Enhancement

Hanwook Chung, *Student Member, IEEE*, Eric Plourde, *Member, IEEE*, and Benoit Champagne, *Senior Member, IEEE*

***Abstract*—In this letter, we introduce a discriminative training algorithm of the basis vectors in the nonnegative matrix factorization (NMF) model for single-channel speech enhancement. 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 consider the probabilistic generative model (PGM) of classification, which is specified by class-conditional densities, along with the NMF model. The update rules of the NMF are jointly obtained with the parameters of the class-conditional densities using the expectation–maximization (EM) algorithm, which guarantees convergence. Experimental results show that the proposed algorithm provides better performance in speech enhancement than the benchmark algorithms.**

***Index Terms*—Classification, discriminative training, expectation–maximization (EM), nonnegative matrix factorization (NMF), probabilistic generative model (PGM), single-channel speech enhancement.**

I. Introduction

NUMEROUS algorithms for single-channel speech enhancement have been proposed such as spectral subtraction [1], minimum mean-square error (MMSE) estimation [2], and subspace decomposition [3]. However, these algorithms use a minimal amount of *a priori* information about the speech and noise and, consequently, tend to provide limited performances, especially when the speech is contaminated by adverse noise such as under low input signal-to-noise ratio (SNR) or nonstationary noise conditions. Recently, the nonnegative matrix factorization (NMF) approach has been successfully applied to various problems, such as source separation [4], speech enhancement [5], and image representation [6]. In general, NMF is a dimensionality reduction tool, which decomposes a given matrix into basis and activation matrices with non-negative elements constraint [7], [8]. In audio and speech applications, the magnitude or power spectrum is interpreted as a linear combination of the basis vectors, which play a key role in the enhancement or separation process.

In a supervised NMF-based framework, the basis vectors are obtained for each source independently during the training stage, and later used during the separation stage. However, one of the main problems is that the basis vectors of the different signal sources may share similar characteristics. For example, the basis vectors of the speech spectrum can represent the noise spectrum and hence, the enhanced speech may contain noise components that have similar features. Considering a specific application, [9] aims to enhance the speech contaminated with highly correlated babble noise, by exploiting a statistical model based on hidden Markov model for the babble noise. More generally, one possible remedy to the aforementioned problem is to use discriminative training criteria, in which the goal is to train the basis vectors of each source in a way that prevents them from representing other sources. In [10], the cross-coherence of the basis vectors is added as a penalty term to the NMF cost function. The authors in [11] and [12] propose to use additional training data which are generated by mixing, e.g., adding or concatenating, the isolated training data of each source. However, these approaches are based on heuristic multiplicative update (MU) rules which do not guarantee the convergence of the NMF in general [8]. Although the NMF update in [13] guarantees local convergence by using a stochastic gradient descent method, the resulting algorithm is limited to pairwise training, i.e., combination of the clean speech with each different type of noise. Moreover, in [12] and [13], the discriminative bases are obtained indirectly by means of the activation matrix estimated from the mixed training data, and hence lacks in a precise interpretation in terms of discrimination.

In this letter, we propose a new algorithm for the discriminative training of the basis vectors in the NMF model. Specifically, the basis vectors for all the clean speech and noise sources are estimated simultaneously during the training stage by incorporating the concept of classification. To this end, we consider the probabilistic generative model (PGM) of classification specified by class-conditional densities [14], [15], along with the NMF model [16]. The main idea is to estimate the basis matrices, during the training stage, by constraining them to belong to one of several speech and noise classes. Within this extended statistical framework, the update rules of the NMF model are jointly obtained along with the parameters of the class-conditional densities via the expectation–maximization (EM) algorithm. Convergence to a stationary point is guaranteed by the properties of the EM algorithm [8], [16]. Experimental results of perceptual evaluation of speech quality (PESQ) [29], source-to-distortion ratio (SDR) [30], and segmental SNR (SSNR) show that the proposed algorithm provides better enhancement performance than the benchmark algorithms.

II. 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 $\mathbf{W} = [w_{km}] \in$

Manuscript received January 19, 2016; accepted February 17, 2016. Date of publication February 23, 2016; date of current version March 16, 2016. This work was supported by Microsemi Corporation (Ottawa, Canada) and NSERC (Govt. of Canada). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Eric Moreau.

H. Chung and B. Champagne are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0G4 Canada (e-mail: hanwook.chung@mail.mcgill.ca; benoit.champagne@mcgill.ca).

E. Plourde is with the Department of Electrical and Computer Engineering, Sherbrooke University, Sherbrooke, QC J1K 2R1 Canada (e-mail: eric.plourde@usherbrooke.ca).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2532903

1070-9908 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

$\mathbb{R}_+^{K\times M}$ is a basis matrix, $\mathbf{H}=[h_{ml}]\in\mathbb{R}_+^{M\times L}$ is an activation matrix, $\mathbb{R}_+$ denotes the set of nonnegative real numbers, and M is the number of basis vectors. The factorization is obtained by minimizing a cost function, such as the Kullback-Leibler (KL) divergence. In this case, the solutions can be obtained iteratively using the following MU rules [7]

$$\mathbf{W}\leftarrow\mathbf{W}\otimes\frac{(\mathbf{V}/(\mathbf{W}\mathbf{H}))\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T},\quad \mathbf{H}\leftarrow\mathbf{H}\otimes\frac{\mathbf{W}^T(\mathbf{V}/(\mathbf{W}\mathbf{H}))}{\mathbf{W}^T\mathbf{1}} \tag{1}$$

where the operation $\otimes$ denotes element-wise multiplication, $/$ and the quotient line are element-wise division, $\mathbf{1}$ is a $K\times L$ matrix with ones, and the superscript T is the matrix transpose.

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., $|Y_{kl}|\approx|S_{kl}|+|N_{kl}|$, where $k\in\{1,\ldots,K\}$ and $l\in\{1,\ldots,L\}$ are the frequency bin and time frame indices [4], [5], [17]. Hence, in this work, $\mathbf{V}=[v_{kl}]$ contains the magnitude spectrum values of either one of the noisy speech, clean speech and noise, as indicated by subscripts or superscripts Y, S, and N, respectively. In a supervised framework, $\mathbf{W}_S$ and $\mathbf{W}_N$ are obtained first during the training stage, by applying (1) to the training data $\mathbf{V}_S$ and $\mathbf{V}_N$. In the enhancement stage, the activation matrix $\mathbf{H}_Y=[\mathbf{H}_S^T\mathbf{H}_N^T]^T$ is estimated by applying the activation update to $\mathbf{V}_Y$, while fixing $\mathbf{W}_Y=[\mathbf{W}_S\mathbf{W}_N]$. Then, the clean speech spectrum can be estimated using a Wiener filter as [8], [17]

$$\hat{S}_{kl}=\frac{\hat{p}_{kl}^S}{\hat{p}_{kl}^S+\hat{p}_{kl}^N}Y_{kl} \tag{2}$$

where $\hat{p}_{kl}^S$ and $\hat{p}_{kl}^N$ 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 [18]

$$\hat{p}_{kl}^S=\tau_S\hat{p}_{k,l-1}^S+(1-\tau_S)([\mathbf{W}_S\mathbf{H}_S]_{kl})^2 \tag{3}$$

$$\hat{p}_{kl}^N=\tau_N\hat{p}_{k,l-1}^N+(1-\tau_N)([\mathbf{W}_N\mathbf{H}_N]_{kl})^2 \tag{4}$$

where τ_S and τ_N are the smoothing factors for the speech and noise, and $[\cdot]_{kl}$ denotes the (k,l)th entry of its matrix argument. The time-domain enhanced speech signal is obtained via inverse STFT followed by the overlap-add method.

III. Probabilistic Generative Models

A. NMF Model With KL-Divergence

In [16], [19], the NMF model with KL-divergence is described within a statistical framework as summarized below. Each entry of a non-negative matrix $\mathbf{V}=[v_{kl}]$ is assumed to be a sum of M latent variables as

$$v_{kl}=\sum_{m=1}^{M}c_{kl}^m. \tag{5}$$

The mth latent variable c_{kl}^m is assumed to be drawn from a Poisson distribution[1] parameterized by w_{km} and h_{ml}

$$p(c_{kl}^m|w_{km}h_{ml})=(w_{km}h_{ml})^{c_{kl}^m}e^{-w_{km}h_{ml}}/(c_{kl}^m!). \tag{6}$$

[1]Note that the approximation of v_{kl} as a sum of integer variables in (5) can be justified by assuming a large dynamic range for the former quantity, which in practice can be realized by a proper scaling of the observations, e.g., magnitude spectra [5], [20].

The maximum likelihood (ML) estimates of the parameters w_{km} and h_{ml}, given the observation v_{kl}, are obtained via the EM algorithm. During the expectation step, the posterior distribution $p(c_{kl}^m|v_{kl})$ is calculated which is shown to follow a binomial distribution. In the maximization step, the parameters are estimated by maximizing the expected complete-data log-likelihood function given, up to a constant term, as

$$\mathcal{L}_C(\mathbf{W},\mathbf{H})\stackrel{c}{=}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(\sum_{m=1}^{M}-w_{km}h_{ml}+\bar{c}_{kl}^m\ln(w_{km}h_{ml})\right) \tag{7}$$

where $\bar{c}_{kl}^m$ is the conditional expectation of the latent variable c_{kl}^m with respect to the posterior distribution, i.e.,

$$\bar{c}_{kl}^m\triangleq E[c_{kl}^m|v_{kl}]=v_{kl}\frac{w_{km}h_{ml}}{\sum_{m'}w_{km'}h_{m'l}}. \tag{8}$$

Iterative NMF solutions obtained through the EM algorithm have similar forms as the MU rules in (1). The scale indeterminacies in w_{km} and h_{ml}, which appear as a product in the distribution in (6), can be prevented by normalizing the columns of $\mathbf{W}$ using the l_1-norm after estimating $\mathbf{W}$, and computing $\mathbf{H}$ accordingly [6], [21], [22].

B. Classification Model

In the classification problem, the input vector $\boldsymbol{w}\in\mathbb{R}^K$ under test is assumed to belong to one of I classes. The goal is to find a corresponding partition of the observation space $\mathbb{R}^K$ (i.e., decision regions) that will minimize the classification errors, by using the training data and their class labels. There are two main approaches to solve this problem: *PGM*, which models the joint distribution of the input data and class labels, and *discriminative modeling*, where the aim is to maximize the posterior class probability (PCP) [14], [15]. In this work, we consider the PGM since it can provide the necessary *a priori* distributions to be used in the proposed discriminative training of the NMF models.[2]

The PGM can be described by a class-conditional density $p(\mathbf{w}|d_i=1)=\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$ [15], where $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $d_i\in\{0,1\}$ is a target class label for the class $i\in\{0,1,\ldots,I-1\}$. For a given training set $\mathbf{W}=[\mathbf{w}_1,\ldots,\mathbf{w}_M]$ and $\mathbf{D}=[\mathbf{d}_1,\ldots,\mathbf{d}_M]$, where $\mathbf{d}_m=[d_{im}]$ is a $I\times 1$ target class label vector with $\sum_i d_{im}=1$, and assuming that the columns $\mathbf{w}_m$ are independently drawn, the likelihood function is given by

$$p(\mathbf{W},\mathbf{D}|\boldsymbol{\theta})=\prod_{m=1}^{M}\prod_{i=0}^{I-1}[p_i\mathcal{N}(\mathbf{w}_m|\boldsymbol{\mu}_i,\boldsymbol{\Sigma})]^{d_{im}} \tag{9}$$

where $\boldsymbol{\theta}=\{\{p_i,\boldsymbol{\mu}_i\}_{i=0}^{I-1},\boldsymbol{\Sigma}\}$ is a PGM parameter set for classification and $p_i\triangleq p(d_i=1)$ is the prior class probability. The set $\boldsymbol{\theta}$ can be simply estimated via the ML criterion. Using Bayes' theorem, the PCP of class i given the observation $\mathbf{w}$ can be shown as

$$p(d_i=1|\mathbf{w})=\frac{p(\mathbf{w}|d_i=1)p_i}{\sum_j p(\mathbf{w}|d_j=1)p_j} \tag{10}$$

which is known as the softmax function [15].

[2]We emphasize that the concept of discriminative training proposed in this work is different from the discriminative modeling in [14] and [15].

IV. Proposed Algorithm

A. Discriminative Training Stage

In [16], a gamma distribution was used for representing the *a priori* information about the basis vectors, as it is shown to be a conjugate prior to the Poisson distribution. Here, we consider instead the Gaussian distribution in order to incorporate PGM-based classification into the proposed framework. We use the class index $i = 0$ for the speech and $i = 1, \ldots, I-1$ for the different noise types. For given training data sets of the clean speech and noises $\mathbf{V} = \{\mathbf{V}^i\}_{i=0}^{I-1}$, our goal is to estimate $\mathbf{W} = \{\mathbf{W}^i\}_{i=0}^{I-1}, \mathbf{H} = \{\mathbf{H}^i\}_{i=0}^{I-1}$ and $\boldsymbol{\theta}$ jointly. For simplicity, we consider a diagonal covariance matrix for the PGM in (9), i.e., $\boldsymbol{\Sigma} = \text{diag}\{\sigma_k^2\}$. Denoting by M_i the number of basis vectors in class i (such that $M = \sum_i M_i$), the likelihood function $p(\mathbf{W}, \mathbf{D}|\boldsymbol{\theta})$ in (9) can be simply rearranged as

$$p(\mathbf{W}|\boldsymbol{\theta}) = \prod_{i=0}^{I-1} \prod_{m=1}^{M_i} p_i \mathcal{N}(\mathbf{w}_m^i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \tag{11}$$

where we omit the dependence on $\mathbf{D}$ hereafter for convenience. For the activations, we employ sparse NMF regularization, which implies that a restricted number of basis vectors will represent the spectrum dominantly. This type of approach is shown to be efficient in the training of the so-called parts-based features of the spectrum [23], as well as for discriminative training [13]. Within the PGM statistical framework, sparsity can be implemented by modeling the entries of $\mathbf{H}$ with an exponential distribution [24]. Assuming that the entries are independent and identically distributed, the prior of $\mathbf{H}$ can be written as

$$p(\mathbf{H}) = \prod_{i=0}^{I-1} \lambda^{M_i L_i} \exp\left(-\lambda \sum_m \sum_l h_{ml}^i\right) \tag{12}$$

where the parameter λ controls the degree of sparsity. The latter is chosen empirically in this work, as usually performed in practical implementations, e.g., [24].

The basis and activation matrices are obtained through a maximum *a posteriori* (MAP) estimator by maximizing the following criterion derived after some simplifications:

$$\ln p(\mathbf{V}, \mathbf{W}, \mathbf{H}|\boldsymbol{\theta}) = \ln p(\mathbf{V}|\mathbf{W}, \mathbf{H}) + \ln p(\mathbf{W}|\boldsymbol{\theta}) + \ln p(\mathbf{H}). \tag{13}$$

Application of the EM algorithm to (13) leads to maximizing the following criterion in its maximization step [8], [15]:

$$\mathcal{L}_C(\mathbf{W}, \mathbf{H}|\boldsymbol{\theta}) = \sum_{i=0}^{I-1} \mathcal{L}_C(\mathbf{W}^i, \mathbf{H}^i) + \ln p(\mathbf{W}|\boldsymbol{\theta}) + \ln p(\mathbf{H}) \tag{14}$$

where $\mathcal{L}_C(\mathbf{W}^i, \mathbf{H}^i)$ is given by (7). The partial derivative of (14) with respect to w_{km}^i can be shown to be

$$\frac{\partial \mathcal{L}_C}{\partial w_{km}^i} = -\sum_{l=1}^{L_i} h_{ml}^i + \frac{1}{w_{km}^i} \sum_{l=1}^{L_i} \overline{c}_{kl}^{m,i} + (\mu_{ik} - w_{km}^i)\sigma_k^{-2} \tag{15}$$

where μ_{ik} denotes the kth entry of the mean vector $\boldsymbol{\mu}_i$, $\overline{c}_{kl}^{m,i}$ is defined as in (8), and L_i is the number of frames in class i. Equation (15) leads to solving the following second-order polynomial equation:

$$\sigma_k^{-2}(w_{km}^i)^2 + \left(\sum_{l=1}^{L_i} h_{ml}^i - \mu_{ik}\sigma_k^{-2}\right) w_{km}^i - \sum_{l=1}^{L_i} \overline{c}_{kl}^{m,i} = 0. \tag{16}$$

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

$$(w_{km}^i)^{(r+1)} = \frac{-q_{i1} + \sqrt{q_{i1}^2 + 4q_0 q_{i2}}}{2q_0} \tag{17}$$

where the superscript (r) denotes the rth iteration, $q_0 = (\sigma_k^{-2})^{(r)}$, $q_{i1} = \sum_l (h_{ml}^i)^{(r)} - \mu_{ik}^{(r)}(\sigma_k^{-2})^{(r)}$, and $q_{i2} = \sum_l (\overline{c}_{kl}^{m,i})^{(r)}$. It is easy to show that (17) takes only positive values. Since σ_k^2 and $\overline{c}_{kl}^{m,i}$ are positive, the product term inside the square root $q_0 q_{i2}$ is also positive, which implies that the numerator in (17) is positive. It is worth noting that employing the Gaussian-distributed *a priori* model can be reasonable, since we can justify $P_r[\mathbf{W} < 0] \approx 0$ for the prior of $\mathbf{W}$ with positive means and small variances such that $\mu_{ik} \gg \sigma_k^2$, which is verified through our experiments. As mentioned in Section III-A, we normalize the columns of $\mathbf{W}$ after computing (17). Following a similar approach as for the basis estimation, the update rule of h_{ml}^i is obtained as

$$(h_{ml}^i)^{(r+1)} = \frac{\sum_k (\overline{c}_{kl}^{m,i})^{(r)}}{\sum_k (w_{kl}^i)^{(r+1)} + \lambda}. \tag{18}$$

The parameter set $\boldsymbol{\theta}$ is estimated by maximizing the following marginal likelihood, where $\mathbf{W}$ is considered as a latent variable [15], [25]

$$p(\mathbf{V}|\mathbf{H}, \boldsymbol{\theta}) = \int p(\mathbf{V}, \mathbf{W}|\mathbf{H}, \boldsymbol{\theta}) d\mathbf{W}. \tag{19}$$

Since this integration cannot be evaluated analytically, we use the Laplace approximation [15], [26]. Accordingly, the log marginal likelihood can be written, up to a constant term, as

$$\ln p(\mathbf{V}|\mathbf{H}, \boldsymbol{\theta}) \overset{c}{\approx} \ln p(\mathbf{V}, \hat{\mathbf{W}}|\mathbf{H}, \boldsymbol{\theta}) - \frac{1}{2} \ln |\mathbf{A}| \tag{20}$$

where $\hat{\mathbf{W}} \triangleq \mathbf{W}^{(r+1)}$ denotes the MAP solution from (17), and $\mathbf{A} \in \mathbb{R}^{KM \times KM}$ is a Hessian matrix. In particular, we can write $|\mathbf{A}| = \prod_{i=1}^{I-1} \prod_{m=1}^{M_i} |\mathbf{A}_{im}|$ where $\mathbf{A}_{im} \in \mathbb{R}^{K \times K}$ is defined as

$$\mathbf{A}_{im} = -\nabla_{\mathbf{w}_m^i} \nabla_{\mathbf{w}_m^i} \ln p(\mathbf{W}^i|\mathbf{V}^i, \mathbf{H}^i, \boldsymbol{\theta})|_{\mathbf{w}_m^i = \hat{\mathbf{w}}_m^i}. \tag{21}$$

For further simplification, we assume that $\mathbf{W}$ is *well determined* which implies that the MAP solution is close to the ML estimate for sufficiently large data set, corresponding to a sharply peaked $p(\mathbf{W}|\mathbf{V}, \mathbf{H}, \boldsymbol{\theta})$ [15], [25]. As a result, the Hessian term in (20) can be neglected[3] since $\partial \ln |\mathbf{A}| / \partial \boldsymbol{\theta} \approx 0$, and estimating $\boldsymbol{\theta}$ becomes equivalent to maximizing (13). The set $\boldsymbol{\theta}$ is then simply found by applying the ML criterion to $p(\hat{\mathbf{W}}|\boldsymbol{\theta})$, where the resulting estimate in a closed form[4] is interleaved with the EM update, as

$$p_i = \frac{M_i}{M}, \mu_{ik}^{(r+1)} = \frac{1}{M_i} \sum_{m=1}^{M_i} (w_{km}^i)^{(r+1)} \tag{22}$$

[3] See [25], [26] and Chap. 12 in [15] for similar applications of this approach.

[4] Compared to using a gamma-distributed PGM as a prior, which requires an iterative process for the hyper-parameter estimation [16], we verified that employing a Gaussian-distributed PGM is more efficient both in terms of computation and enhancement performance.

$$(\sigma_k^2)^{(r+1)} = \frac{1}{M}\sum_{i=0}^{I-1}\sum_{m=1}^{M_i}\left[(w_{km}^i)^{(r+1)} - \mu_{ik}^{(r+1)}\right]^2. \quad (23)$$

As for the initialization, we generate positive random numbers for $\mathbf{W}^i$ and $\mathbf{H}^i$, and subsequently apply (22)–(23) for $\boldsymbol{\theta}$. The proposed discriminative NMF algorithm with class probabilities will be referred to as DCP.

B. Enhancement Stage

The enhancement stage is similar to the standard method described in Section II. Upon fixing $\mathbf{W}_Y = [\mathbf{W}_S \mathbf{W}_N]$, the activation matrix $\mathbf{H}_Y = [\mathbf{H}_S^T \ \mathbf{H}_N^T]^T$ is estimated by applying the activation update rule to $\mathbf{V}_Y$ with sparse regularization as in (18). Note that, since $\mathbf{W}_Y$ is fixed, we can simply use the activation update given in (1) by adding $\lambda\mathbf{1}$ to the denominator. Moreover, $\boldsymbol{\theta}$ can be used for the noise classification based on (10) in advance to the enhancement. In this case, the additional noise basis vector $\mathbf{w}$ needed for the purpose of classification is obtained through $[\mathbf{W}_S \ \mathbf{w}]$ by applying (1) to $\mathbf{V}_Y$. However, we simply assume that the noise type is known in this letter.

V. Experiments

We used clean speech from the TSP database [27] and noise from the NOISEX database [28], where the sampling rate of all signals was set 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 ($i=0$), 20 speakers (10 males and 10 females) were considered, whereas the factory 1 ($i=1$), buccaneer 1 ($i=2$), and HF Channel ($i=3$) noises were selected. We examined both the speaker-dependent (SD) and speaker-independent (SI) cases, i.e., one basis matrix per speaker for the SD and one universal basis matrix covering all speakers for the SI. For the SD, the training data consisted of 20 sentences (45 s) for each speaker, whereas for the SI, one sentence from each speaker was selected for a total of 20 sentences (50 s). For each noise type, 30-s samples were used as the training data. The test speech signals consisted of two sentences (6 s) which were not included in the training data. The noisy speech was generated from the test signals by adding the noise to the clean speech to obtain input SNR of 0, 5, and 10 dB. We used $M_i = 80$ basis vectors for all $i \in \{0,1,2,3\}$. Sparsity and temporal smoothing factors were selected as $\lambda = 0.5$ and $(\tau_S, \tau_N) = (0.4, 0.9)$.

Fig. 1 shows the PCPs of the basis vectors for the speech class $i = 0$, i.e., $p(d_0 = 1|\mathbf{w}_m)$, computed by (10) based on the estimated set $\boldsymbol{\theta}$. The interval $1 \le m \le 80$ corresponds to the universal speech basis vectors where the PCPs should be close to one, whereas the interval $81 \le m \le 320$ corresponds to the noise basis vectors, i.e., $1 \le i \le 3$ (80 vectors for each i), where the PCPs should be close to zero. We can see that the basis vectors obtained from the DCP method lead to a more precise classification. In turn, this implies that the basis vectors of each source will be less likely to represent each other. Similar results were found for the other classes and even better results were observed for the SD basis vectors.

We used PESQ [29], SDR [30], and SSNR as the objective measures, where a higher value indicates a better result. To compare the proposed approach, we implemented the standard NMF method described in Section II, and several discriminative NMF (DNMF) algorithms, where we will refer to each algorithm using its reference number. We used $\alpha = 1, \lambda = 200$ for [10], $\lambda = 0.01$ for [11], and $\mu = 0.1$, KL objective for [12] (see the references for the parameter description). Note that [10] and [12] are based on a pairwise training, where the bases are estimated from $\mathbf{V} = \{\mathbf{V}_0, \mathbf{V}_i\}$ for each $i = \{1,2,3\}$, i.e., different $\mathbf{W}_S$ for each noise type. On the contrary, [11] and the DCP method estimate all basis matrices simultaneously from $\mathbf{V} = \{\mathbf{V}_i\}_{i=0}^3$. Tables I and II show the average results of the buccaneer 1 and HF Channel noises for the SD and SI. We can see that the best values were obtained with the DCP method. Similar results were observed for the factory 1 noise and 10-dB input SNR.

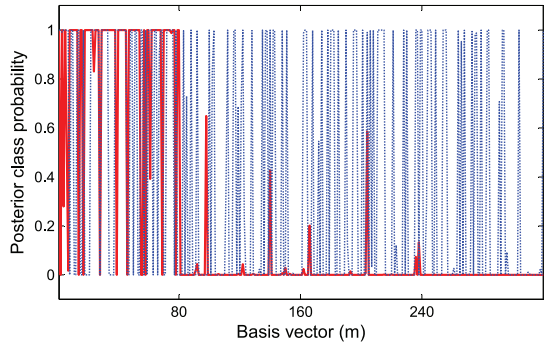

Fig. 1. Posterior class probabilities $p(d_0 = 1|\mathbf{w}_m)$. Red-solid indicates the proposed DCP method, whereas blue-dashed denotes the standard NMF.

TABLE I
AVERAGE RESULTS FOR BUCCANEER 1 AND HF CHANNEL NOISES (SD)

Input SNR		Eval.	Noisy	NMF	DNMF [10]	DNMF [11]	DNMF [12]	DCP
Buccaneer 1	0 dB	PESQ	1.20	1.71	1.83	1.92	2.01	**2.17**
		SDR	0.05	5.06	6.20	6.88	7.19	**8.39**
		SSNR	−4.54	−1.10	−0.36	0.79	1.07	**1.93**
	5 dB	PESQ	1.54	2.09	2.21	2.25	2.24	**2.41**
		SDR	5.04	9.71	10.37	10.28	10.57	**11.67**
		SSNR	−1.37	2.15	2.59	3.43	3.57	**4.48**
HF Channel	0 dB	PESQ	1.17	1.71	1.90	1.93	2.01	**2.25**
		SDR	0.06	6.31	7.49	8.58	8.53	**9.64**
		SSNR	−4.53	−0.34	0.60	1.87	1.79	**3.79**
	5 dB	PESQ	1.44	2.06	2.25	2.22	2.21	**2.47**
		SDR	5.04	10.75	11.35	11.68	11.78	**12.54**
		SSNR	−1.36	2.88	3.56	4.42	4.41	**6.05**

TABLE II
AVERAGE RESULTS FOR BUCCANEER 1 AND HF CHANNEL NOISES (SI)

Input SNR		Eval.	Noisy	NMF	DNMF [10]	DNMF [11]	DNMF [12]	DCP
Buccaneer 1	0 dB	PESQ	1.20	1.64	1.66	1.85	1.85	**2.05**
		SDR	0.05	4.38	4.88	6.06	5.79	**7.47**
		SSNR	−4.54	−1.51	−1.24	0.25	0.35	**0.99**
	5 dB	PESQ	1.54	2.02	2.04	2.17	2.06	**2.33**
		SDR	5.04	9.13	9.60	9.42	8.97	**11.03**
		SSNR	−1.37	1.69	1.92	2.88	2.73	**3.66**
HF Channel	0 dB	PESQ	1.17	1.64	1.72	1.80	1.91	**2.17**
		SDR	0.06	5.70	6.18	7.81	7.92	**9.27**
		SSNR	−4.53	−0.75	−0.39	1.41	1.56	**3.08**
	5 dB	PESQ	1.44	2.00	2.09	2.10	2.14	**2.43**
		SDR	5.04	10.25	10.61	10.90	11.01	**12.30**
		SSNR	−1.36	2.47	2.78	3.94	4.09	**5.44**

VI. Conclusion

A discriminative training algorithm of the basis vectors in the NMF model for single-channel speech enhancement has been proposed. The basis matrices of the clean speech and noises were estimated during the training stage by constraining them to belong to different classes. To this end, we considered the PGM with class-conditional densities along with the NMF model. The update rules of the extended NMF model were jointly obtained with PGM parameters via the EM algorithm. Experiments showed that the proposed algorithm provided better results than the benchmark algorithms.

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] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1614–1623, Nov. 2008.

[3] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness constraint," *IEEE Trans. Audio Speech Lang. 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 Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[6] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 683–695, May 2006.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2001, pp. 556–562.

[8] C. Fevotte, 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.

[9] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 5, pp. 998–1011, May 2013.

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

[11] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 3749–3753.

[12] F. Weninger, J. L. Roux, J. R. Hershy, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Proc. Interspeech*, Sep. 2014, pp. 865–869.

[13] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-euclidean sparse NMF via bilevel optimization with applications to speech enhancement," in *Proc. 4th Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, May 2014, pp. 1–5.

[14] I. Ulusoy and C. M. Bishop, "Generative versus discriminative models for object recognition," in *Proc. Comput. Vision Pattern Recog.*, Jun. 2005, pp. 258–265.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[16] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, vol. 2009, no. 4, pp. 1–17, 2009.

[17] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proc. Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2011, pp. 45–48.

[18] 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.

[19] W. Buntine and A. Jakulin, "Discrete component analysis," in *Proc. Subspace Latent Struct. Feature Selection*, 2006, pp. 1–33.

[20] M. D. Hoffman, "Poisson-uniform nonnegative matrix factorization," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 5361–5364.

[21] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2006, pp. 621–624.

[22] M. Sun, Y. Li, F. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[23] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraint," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Nov. 2004.

[24] M. N. Schmidt and J. Larsen, "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *Proc. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2008, pp. 486–491.

[25] C. M. Bishop, "Bayesian PCA," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 1998, pp. 382–388.

[26] O. Dikmen and C. Fevotte, "Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma–Poisson model," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5163–5175, Oct. 2012.

[27] P. Kabal, "TSP speech database," McGill Univ., Montreal, QC, Canada, Tech. Rep. 09-02, 2002.

[28] 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 Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[29] 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.

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