A SUPERVISED MULTI-CHANNEL SPEECH ENHANCEMENT ALGORITHM BASED ON BAYESIAN NMF MODEL

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

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

ABSTRACT

In this paper, we introduce a supervised multi-channel speech enhancement algorithm based on a Bayesian multi-channel non-negative matrix factorization (MNMF) model. In the proposed framework, we consider the probabilistic generative model (PGM) of MNMF, specified by Poisson-distributed latent variables and gamma-distributed priors. In the training stage, the MNMF parameters of the speech and noise sources are estimated via the variational Bayesian expectation-maximization (VBEM) algorithm. In the enhancement stage, the clean speech signal is estimated via the MNMF-based minimum variance distortionless response (MVDR) beamformer. To further improve the enhanced speech quality, we efficiently combine the MNMF-based beamforming technique with a classical unsupervised single-channel enhancement method. Experiments show that the proposed method can provide better enhancement performance than the selected benchmarks.

***Index Terms*—** Multi-channel speech enhancement, MVDR beamforming, non-negative matrix factorization, probabilistic generative model, variational Bayesian expectation-maximization

1. INTRODUCTION

Numerous single and multi-channel speech enhancement algorithms have been proposed in the past. They aim to remove the background noise from a noisy speech signal, in order to improve its quality or intelligibility, and find diverse applications including mobile telephony, hearing aid and speech recognition. Compared to the single-channel algorithms, the main advantage of the multi-channel algorithms is that they can exploit the spatial features of the acoustic field through a beamformer, which can be designed to extract the clean speech from a given direction in an optimal way. In this context, classical beamforming techniques incldue: minimum variance distortionless response (MVDR) or linearly constrained minimum variance (LCMV) [1, 2], generalized sidelobe canceller (GSC) [3], and eigen-space beamformer [4]. However, these classical methods were originally introduced by using a minimal amount of *a priori* information about the speech and noise and hence, tend to provide limited performance under adverse noise conditions.

Machine learning techniques have been applied to the speech enhancement task in recent years, to better exploit *a priori* information. Among these, the non-negative matrix factorization (NMF) method, which decomposes a given matrix into basis and activation matrices with non-negative elements, has received great attention [5, 6, 7]. The concept of NMF has been extended to the factorization of a tensor into mixing, basis and activation matrices, referred to as multi-channel NMF (MNMF) [8]. In this approach, both the multiplicative update (MU) rules and expectation-maximization (EM) algorithms have been derived to estimate the MNMF parameters, based on the Itakura-Saito (IS) divergence. To better exploit the spatial properties of the sources, references [9, 10] aim at factorizing the spatial covariance matrix (SCM) of the observation in each frequency bin, which is specified by the channel covariance matrices of the individual sources. In [11], the complex-valued Gaussian-distributed latent variables of MNMF are modeled by auto-regressive moving average (ARMA) processes, to better handle the reverberation effects. The MNMF method has shown promising results in multi-channel source separation and speech enhancement. Still, besides the need to improve the enhancement/separation performance, computational complexity remains a critical issue when implementing the MNMF algorithms. That is, compared to the MU-based algorithms, the computational cost increases rapidly as the numbers of NMF basis vectors and microphones increase.

In this paper, we propose a supervised multi-channel speech enhancement algorithm based on a Bayesian MNMF model. Our main contribution is to extend the probabilistic generative model (PGM) of NMF, specified by Poisson-distributed latent variables and gamma-distributed priors [12], to a multi-channel framework. In the training stage, the MNMF parameters of different sources are estimated via the variational Bayesian expectation-maximization (VBEM) algorithm [13]. The main advantage of using the Poisson-distributed PGM, compared to the complex-valued Gaussian-distributed PGM, is the reduced complexity while implementing the VBEM algorithm, since we only need the marginal statistics, as pointed out in [12]. In the enhancement stage, the clean speech signal is estimated via the MNMF-based MVDR beamforming technique. To further improve the enhanced speech quality, we combine the MNMF-based beamforming technique with a classical unsupervised single-channel enhancement method. Experiments show that the proposed method can provide better performance than the selected benchmarks.

Throughout the paper, we use the superscripts T, H and $*$ to denote matrix transpose, Hermitian transpose and complex conjugate operation. The symbols $\mathbb{R}_+$ and $\mathbb{C}$ denote the sets of non-negative real and complex numbers, ! indicates factorial, $[\cdot]_{ij}$ denotes the (i, j)-th entry of its matrix argument, $|\cdot|$ indicates the (element-wise) magnitude computation, $\mathbf{I}_J$ indicates the $J \times J$ identity matrix, and $\propto$ denotes linear proportionality. The imaginary unit is expressed by $\jmath = \sqrt{-1}$, while $\angle Y$ represents the phase of a complex number Y.

2. MNMF-BASED SPEECH ENHANCEMENT

Assuming that the length of a mixing filter, i.e., the acoustic impulse response, is shorter than the short-time Fourier transform (STFT)

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

978-1-7281-1295-4/18/$31.00 ©2018 IEEE

analysis window length, the multi-channel convolutive noisy speech signal can be expressed in the STFT domain as [8, 10]

$$Y^j_{kl} = Z^j_{S,kl} + Z^j_{N,kl} = \tilde{A}^S_{kj} S_{kl} + Z^j_{N,kl} \tag{1}$$

where Y^j_{kl}, $Z^j_{S,kl}$ and $Z^j_{N,kl}$ respectively denote the STFT coefficients of the convolutive noisy speech, clean speech and noise signals at the frequency bin $k \in \{1, ..., K\}$, time frame $l \in \{1, ..., L\}$ and microphone index $j \in \{1, ..., J\}$, $\tilde{A}^S_{kj}$ is the acoustic transfer function (ATF) for the clean speech, and S_{kl} is the STFT of the clean speech point source signal.

For a given tensor $\mathbf{V} = [v^j_{kl}] \in \mathbb{R}^{K \times L \times J}_+$, the MNMF algorithm aims at factorizing it into a mixing matrix $\mathbf{A} = [a_{kj}] \in \mathbb{R}^{K \times J}_+$, a basis matrix $\mathbf{W} = [w_{km}] \in \mathbb{R}^{K \times M}_+$ and an activation matrix $\mathbf{H} = [h_{ml}] \in \mathbb{R}^{M \times L}_+$, where M is the number of basis vectors. Specifically, it seeks to represent each entry of $\mathbf{V}$ in the form of [8]

$$v^j_{kl} \approx \hat{v}^j_{kl} = a_{kj} \sum_{m=1}^{M} w_{km} h_{ml}. \tag{2}$$

The factorization is obtained by minimizing a cost function, such as the Kullback-Leibler (KL) divergence. In audio and speech applications, the KL-based MNMF algorithm is typically applied to a magnitude spectrum-based tensor, e.g., $v^j_{kl} = |Z^j_{S,kl}|$. Hence, w_{km} and h_{ml} become related to the point source spectrum (e.g., $|S_{kl}| = \sum_m w_{km} h_{ml}$), while a_{kj} corresponds to the magnitude value of the ATF (e.g., $a_{kj} = |\tilde{A}^S_{kj}|$). At this point, we further assume that the noise spectrum in (1) can be expressed as $Z^j_{N,kl} = \tilde{A}^N_{kj} N_{kl}$. Although this model is theoretically valid only for a noise signal generated by a point source, it has been widely considered in practice in the MNMF-based approaches, e.g., [10], mainly due to an efficient representation of the signal by means of the basis vectors. Moreover, this enables a possible post-processing of the beamformer output.

A supervised MNMF-based multi-channel speech enhancement framework consists of two stages. In the training stage, the basis matrices of the clean speech and noise, $\mathbf{W}_S = [w^S_{km}] \in \mathbb{R}^{K \times M_S}_+$ and $\mathbf{W}_N = [w^N_{km}] \in \mathbb{R}^{K \times M_N}_+$, are obtained from the training data. In the enhancement stage, by fixing $[\mathbf{W}_S \; \mathbf{W}_N]$, we estimate the mixing and activation matrices of the clean speech and noise from the noisy speech. In this paper, we consider a buffer processing to handle the case of time-varying ATFs as well as to alleviate the so-called over-complete condition. That is, we aim at factorizing $\mathbf{V}^Y_{l_b} = |\mathbf{Y}_{l_b}| \in \mathbb{R}^{K \times L_b \times J}_+$, where $l_b = 1, 2, ...$ is the buffer frame index, L_b is the buffer size and $\mathbf{Y}_{l_b}$ is the noisy speech matrix consisting of the time frames $l \in \{(l_b - 1)L_b + 1, ..., l_b L_b\}$.

Once we obtain the MNMF parameters for a given buffer frame index, the clean speech point source spectrum can be estimated via MVDR beamforming as [14]

$$\hat{S}^B_{kl} = \left(\frac{(\mathbf{R}^N_{kl} + \lambda \mathbf{I}_J)^{-1} \mathbf{b}_k}{\mathbf{b}^H_k (\mathbf{R}^N_{kl} + \lambda \mathbf{I}_J)^{-1} \mathbf{b}_k} \right)^H \mathbf{Y}_{kl} = \hat{\mathbf{G}}^H_{kl} \mathbf{Y}_{kl} \tag{3}$$

where $\lambda > 0$ is the diagonal loading factor, $\mathbf{b}_k = [b^j_k] \in \mathbb{C}^J$ is the steering vector, which is specified by the direction-of-arrival (DoA) under the far field assumption, and $\mathbf{R}^N_{kl} \in \mathbb{C}^{J \times J}$ is the noise correlation matrix. The latter is obtained via temporal smoothing of the initial estimate of $Z^j_{N,kl}$ [15], which can be expressed in terms of the MNMF parameters as

$$[\mathbf{R}^N_{kl}]_{ab} = \tau_C [\mathbf{R}^N_{k,l-1}]_{ab} + (1 - \tau_C) \hat{A}^N_{ka} \left(\hat{A}^N_{kb} \right)^* [\mathbf{W}_N \hat{\mathbf{H}}_N]^2_{kl} \tag{4}$$

where $0 < \tau_C < 1$ is the smoothing constant, $\hat{\mathbf{H}}_N$ is the estimated activation matrix of the noise, $\hat{A}^N_{kj} = \hat{a}^N_{kj} \exp(\jmath \angle \hat{A}^N_{kj})$ is the estimated complex-valued ATF for the noise, and $a, b \in \{1, ..., J\}$. The phase components can be obtained from the noisy speech phase. Finally, the time-domain enhanced speech signal is obtained via inverse STFT, followed by the overlap-add method.

3. PROPOSED TRAINING STAGE

In this section, we first introduce the PGM of MNMF that corresponds to the KL-divergence in a statistical framework. Then, we derive the VBEM algorithm to estimate the MNMF parameters.

3.1. Probabilistic generative model of MNMF

From a statistical perspective, each entry of $\mathbf{V}$ can be constructed as a sum of M latent variables, i.e., $v^j_{kl} = \sum_{m=1}^{M} c^{m,j}_{kl}$. According to [12], the m-th latent variable $c^{m,j}_{kl}$ can be assumed to be drawn from a Poisson distribution parameterized by a^j_k, w_{km} and h_{ml}. That is:

$$p(c^{m,j}_{kl} | a_{kj}, w_{km}, h_{ml}) = \mathcal{P}(c^{m,j}_{kl} | a_{kj} w_{km} h_{ml}) \tag{5}$$

where $\mathcal{P}(x|u) = u^x \exp(-u)/(x!)$ is the Poisson distribution with mean u. Assuming that v^j_{kl} are independently drawn, the log-likelihood function (LLF) of $\mathbf{V}$ can be written as[1]

$$\ln p(\mathbf{V}|\mathbf{A}, \mathbf{W}, \mathbf{H}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} \left(v^j_{kl} \ln \hat{v}^j_{kl} - \hat{v}^j_{kl} - \ln v^j_{kl}! \right). \tag{6}$$

It can be shown that maximizing the LLF with respect to the MNMF parameters is equivalent to minimizing the KL-divergence.

Regarding the prior distributions for $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$, we consider the gamma distribution, which is shown to be the conjugate prior to the Poisson distribution, as given by [5, 12]

$$p(a_{kj}; \alpha_a, \beta_a) = \mathcal{G}(a_{kj}; \alpha_a, \beta_a/\alpha_a) \tag{7}$$

$$p(w_{km}; \alpha_w, \beta_w) = \mathcal{G}(w_{km}; \alpha_w, \beta_w/\alpha_w) \tag{8}$$

$$p(h_{ml}; \alpha_h, \beta_h) = \mathcal{G}(h_{ml}; \alpha_h, \beta_h/\alpha_h) \tag{9}$$

where $\mathcal{G}(x; a, b) = x^{a-1} b^{-a} \exp(-x/b)/\Gamma(a)$ is the gamma distribution with mean ab, $\Gamma(\cdot)$ is the gamma function, and a and b are referred to as the shape and scale parameters, respectively. We consider constant values for the hyper-parameters for each type of matrix factor to avoid over-fitting [5, 12]. Moreover, we assume that the entries of $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$ are independently distributed.

3.2. VBEM algorithm

In many applications of the EM algorithm, evaluating the posterior distribution or computing the expectations with respect to this distribution is analytically intractable. The VBEM algorithm overcomes this difficulty by computing an analytical and efficient approximation to the posterior distribution, and also provides an effective estimation of the hyper-parameters.

Let us denote by $\boldsymbol{\theta}_L = \{\mathbf{C}, \mathbf{A}, \mathbf{W}, \mathbf{H}\}$ the set of latent variables, and by $\boldsymbol{\theta}_R = \{\alpha_a, \beta_a, \alpha_w, \beta_w, \alpha_h, \beta_h\}$ the set of hyper-parameters. For a given observation, we estimate $\boldsymbol{\theta}_L$ and $\boldsymbol{\theta}_R$ via the VBEM algorithm, which consists of two steps at each iteration. In the expectation-step (E-step), we estimate the variational distribution $q(\boldsymbol{\theta}_L)$ that approximates the exact posterior distribution $p(\boldsymbol{\theta}_L | \mathbf{V}; \boldsymbol{\theta}_R)$. In the maximization-step (M-step), we estimate the set $\boldsymbol{\theta}_R$ by maximizing the expectation of the complete-data LLF with respect to the variational distribution, i.e., $\mathbb{E}_{q(\boldsymbol{\theta}_L)}[\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)]$. Additional details are given below.

1) E-step: Based on the mean-field approximation [13], the local optimal solutions can be expressed as

$$q(\mathbf{C})^{(r+1)} \propto \exp\Big(\mathbb{E}_{q(\mathbf{A})^{(r)} q(\mathbf{W})^{(r)} q(\mathbf{H})^{(r)}} [\ln p(\mathbf{V}, \boldsymbol{\theta}_L; \boldsymbol{\theta}_R)] \Big) \tag{10}$$

[1] We note that the sum of independent Poisson random variables x_m with means μ_m is another Poisson random variable with mean $\sum_m \mu_m$.

Table 1. Variational distribution parameters (E-step) and update formula (M-step)

	$\bar{\alpha}$ (E-step)	$\bar{\beta}$ (E-step)	α_q (M-step)	β (M-step)
a_{kj}	$\alpha_a+\sum_k\sum_j\mathbb{E}_q[c_{kl}^{m,j}]$	$\alpha_a/\beta_a+\sum_k\sum_l\mathbb{E}_q[w_{km}]\mathbb{E}_q[h_{ml}]$	$\sum_k\sum_j(\mathbb{E}_q[a_{kj}]/\beta_a-\mathbb{E}_q[\ln a_{kj}]+\ln\beta_a)/(KJ)$	$\sum_k\sum_j\mathbb{E}_q[a_{kj}]/(KJ)$
w_{km}	$\alpha_w+\sum_k\sum_m\mathbb{E}_q[c_{kl}^{m,j}]$	$\alpha_w/\beta_w+\sum_k\sum_m\mathbb{E}_q[a_{kj}]\mathbb{E}_q[h_{ml}]$	$\sum_k\sum_m(\mathbb{E}_q[w_{km}]/\beta_w-\mathbb{E}_q[\ln w_{km}]+\ln\beta_w)/(KM)$	$\sum_k\sum_m\mathbb{E}_q[w_{km}]/(KM)$
h_{ml}	$\alpha_h+\sum_m\sum_l\mathbb{E}_q[c_{kl}^{m,j}]$	$\alpha_h/\beta_h+\sum_m\sum_l\mathbb{E}_q[a_{kj}]\mathbb{E}_q[w_{km}]$	$\sum_m\sum_l(\mathbb{E}_q[h_{ml}]/\beta_h-\mathbb{E}_q[\ln h_{ml}]+\ln\beta_h)/(ML)$	$\sum_m\sum_l\mathbb{E}_q[h_{ml}]/(ML)$

$$q(\mathbf{A})^{(r+1)}\propto\exp\Big(\mathbb{E}_{q(\mathbf{C})^{(r+1)}q(\mathbf{W})^{(r)}q(\mathbf{H})^{(r)}}[\ln p(\mathbf{V},\boldsymbol{\theta}_L;\boldsymbol{\theta}_R)]\Big)$$
$$q(\mathbf{W})^{(r+1)}\propto\exp\Big(\mathbb{E}_{q(\mathbf{C})^{(r+1)}q(\mathbf{A})^{(r+1)}q(\mathbf{H})^{(r)}}[\ln p(\mathbf{V},\boldsymbol{\theta}_L;\boldsymbol{\theta}_R)]\Big)$$
$$q(\mathbf{H})^{(r+1)}\propto\exp\Big(\mathbb{E}_{q(\mathbf{C})^{(r+1)}q(\mathbf{W})^{(r+1)}q(\mathbf{W})^{(r+1)}}[\ln p(\mathbf{V},\boldsymbol{\theta}_L;\boldsymbol{\theta}_R)]\Big)$$

where the superscript (r) denotes the r-th iteration. For convenience, we shall omit (r) and also drop the latent variables appearing as argument of $q(\cdot)$ of the expectation operators, e.g., $\mathbb{E}_{q(w_{km})^{(r)}}[w_{km}]=\mathbb{E}_q[w_{km}]$.

By developing (10), the distribution $q(\mathbf{c}_{kl}^j)$ is shown to be multinomial [12], i.e., $q(\mathbf{c}_{kl}^j)=\mathcal{M}(\mathbf{c}_{kl}^j;v_{kl}^j,\bar{\mathbf{p}}_{kl}^j)$ where[2] $\mathbf{c}_{kl}^j=[c_{kl}^{1,j},...,c_{kl}^{M,j}]$, and the entries of $\bar{\mathbf{p}}_{kl}^j$ are given by

$$\bar{p}_{kl}^{m,j}=\frac{\exp\big(\mathbb{E}_q[\ln a_{kj}]+\mathbb{E}_q[\ln w_{km}]+\mathbb{E}_q[\ln h_{ml}]\big)}{\sum_{m=1}^{M}\exp\big(\mathbb{E}_q[\ln a_{kj}]+\mathbb{E}_q[\ln w_{km}]+\mathbb{E}_q[\ln h_{ml}]\big)}. \quad (11)$$

Accordingly, we have that $\mathbb{E}_q[c_{kl}^{m,j}]=v_{kl}^j\bar{p}_{kl}^{m,j}$. In a similar way, the distributions $q(a_{kj})$, $q(w_{km})$ and $q(h_{ml})$ are shown to be gamma $q(x)=\mathcal{G}(x;\bar{\alpha},\bar{\beta})$, where the values $\bar{\alpha}$ and $\bar{\beta}$ for each matrix factor are summarized in Table 1. For these gamma distributions, we find that $\mathbb{E}_q[\ln x]=\Psi(\bar{\alpha})+\ln\bar{\beta}$ and $\mathbb{E}_q[x]=\bar{\alpha}\bar{\beta}$, where $\Psi(x)\triangleq d\ln\Gamma(x)/dx$.

2) M-step: The shape parameters of the prior gamma distributions in (7)-(9) are obtained based on the Newton's iterative method as

$$\alpha^{(i+1)}=\alpha^{(i)}-\frac{\ln\alpha^{(i)}-\Psi(\alpha^{(i)})+1-\alpha_q}{1/\alpha^{(i)}-\Psi'(\alpha^{(i)})} \quad (12)$$

where $\Psi'(x)=d\Psi(x)/dx$, and i is the iteration index. Detailed formulae for α_q in (12) and β in (7)-(9) for each matrix factor are provided in Table 1.

For initialization, we first generate positive random numbers and subsequently apply the KL-based MU rules [16] to $\bar{\mathbf{V}}=[\sum_{j=1}^{J}v_{kl}^j/J]\in\mathbb{R}_+^{K\times L}$ for several iterations (i.e., 10) [8]. The resulting $\mathbf{W}$ and $\mathbf{H}$ are used as the initial values for $\mathbb{E}_q[w_{km}]$, $\exp(\mathbb{E}_q[\ln w_{km}])$, $\mathbb{E}_q[h_{ml}]$ and $\exp(\mathbb{E}_q[\ln h_{ml}])$. The statistics $\mathbb{E}_q[a_{kj}]$ and $\exp(\mathbb{E}_q[\ln a_{kj}])$ are initialized to 1. The hyperparameters are initialized as $\alpha_a=\alpha_w=\alpha_h=0.001$ and $\beta_a=\beta_w=\beta_h=10$. We use 200 iterations for the VBEM algorithm, whereas 5 iterations are used for estimating the shape parameter in (12). To avoid scale indeterminacies in a_{kj}, w_{km} and h_{ml} which appear as a product in the distribution (5), we include a normalization step at each iteration. That is, motivated by [17], we normalize $\mathbb{E}_q[a_{kj}]$ and $\exp(\mathbb{E}_q[\ln a_{kj}])$ after inferring $q(a_{kj})$, such that $\sum_j\mathbb{E}_q[a_{kj}]=1$ and $\sum_j\exp(\mathbb{E}_q[\ln a_{kj}])=1$. Also, we normalize $\mathbb{E}_q[w_{km}]$ and $\exp(\mathbb{E}_q[\ln w_{km}])$ after inferring $q(w_{km})$, such that $\sum_k\mathbb{E}_q[w_{km}]=1$ and $\sum_k\exp(\mathbb{E}_q[\ln w_{km}])=1$.

4. PROPOSED ENHANCEMENT STAGE

In the enhancement stage, for each buffer, we first estimate the mixing elements (i.e., a_{kj}^S, a_{kj}^N) and activation elements (i.e., h_{ml}^S, h_{ml}^N) elements from the noisy speech via the VBEM algorithm. Subsequently, we compute the noise correlation matrix based on (4). Specifically, we replace $|\hat{A}_{kj}^N|$, w_{km}^N and h_{ml}^N with $\mathbb{E}_q[a_{kj}^N]$, $\mathbb{E}_q[w_{km}^N]$ and $\mathbb{E}_q[h_{ml}^N]$, respectively [7]. The DoA of the speech source, which is needed to compute the steering vectors $\mathbf{b}_k$, is estimated from the noisy speech. Several source localization methods have been proposed, such as the angular spectrum-based and clustering-based methods (see [18] for detailed discussion). In this paper, we implement the method introduced in [19]. The clean speech spectrum then can be estimated via MNMF-based MVDR beamforming given by (3).

To further reduce the residual noise components, the application of a single-channel enhancement algorithm to the beamformer output as a post-processor has been proposed, e.g., [20]. However, such post-processing may introduce clean speech distortion as well. Besides, the combination of a classical unsupervised method with a NMF-based method has been proposed in a single-channel speech enhancement task, e.g., [21]. We adopt this approach to compensate for the clean speech distortion. Let $\bar{\mathbf{Y}}_{kl}\in\mathbb{C}^J$ denote the preprocessed noisy speech spectrum obtained by applying a classical single-channel algorithm to each channel of the noisy speech. We estimate the magnitude components of the clean speech spectrum via the geometric mean (GM) of the magnitude spectra of i) the output of the MVDR beamformer applied to $\mathbf{Y}_{kl}$, ii) same applied to $\bar{\mathbf{Y}}_{kl}$, and ii) the post-processed MVDR output based on NMF-based Wiener filtering [7]. By taking the phase from $\hat{S}_{kl}^B$, the proposed enhanced speech spectrum can be written as

$$\hat{S}_{kl}^G=|\hat{S}_{kl}^B|^{1/3}\,|\hat{\mathbf{G}}_{kl}^H\bar{\mathbf{Y}}_{kl}|^{1/3}\,\left|\frac{\hat{p}_{kl}^S}{\hat{p}_{kl}^S+\hat{p}_{kl}^N}\hat{S}_{kl}^B\right|^{1/3}e^{\jmath\angle\hat{S}_{kl}^B} \quad (13)$$

where $\hat{p}_{kl}^S$ and $\hat{p}_{kl}^N$ are the power spectral densities (PSD) of the clean speech and noise obtained via temporal smoothing of the NMF-based periodograms [7]. The proposed method, i.e., variational inference on the MNMF model, will be referred to as VMNMF.

5. EXPERIMENTS

In this section, after describing the data sets and general methodology, we present and discuss the experimental results.

5.1. Data set

We conducted the experiments using the 4-th CHiME challenge corpus [22], where the sampling rate of all signals was set to 16 kHz. The speech and noise files were divided into two disjoint groups: i) *training data*, used for estimating the basis matrix in the training stage, and ii) *test data*, used in the enhancement stage to evaluate the enhancement performance. The clean speech training data of the CHiME database consists of 101 speakers. We considered a speaker-independent (SI) application, where one *universal* basis matrix covering all speakers is estimated during the training stage. To this end, we randomly selected 40 utterances from each speaker and concatenated them to construct the clean speech training data, resulting in an 8 hours long signal. Regarding the noise training data, we considered the Bus, Cafe, Pedestrian, and Street noises, where we used a 2 hours long signal for each noise type.

Regarding the test data, we used the one referred to as "simulated development data" from the CHiME corpus, which consists

[2] $\mathcal{M}(\mathbf{c};v,\bar{\mathbf{p}})=v!\prod_m(\bar{p}_m)^{c_m}/(c_m!)$, such that $v=\sum_m c_m$.

Table 2. Average results for Bus noise

Input SNR	Eval.	STSA	MNMF -IS-MU	NTF -KL-MU	NTF -IS-MU	VMNMF
-5 dB	PESQ	2.18	2.33	2.25	2.24	**2.38**
	SDR	2.48	8.03	7.34	6.57	**9.47**
	ESTOI	0.57	0.66	0.63	0.61	**0.68**
0 dB	PESQ	2.51	2.64	2.58	2.56	**2.71**
	SDR	6.57	11.43	11.00	10.06	**12.72**
	ESTOI	0.71	0.78	0.77	0.74	**0.80**
5 dB	PESQ	2.80	2.94	2.90	2.87	**3.01**
	SDR	10.28	14.33	14.00	12.81	**15.25**
	ESTOI	0.82	0.87	0.87	0.84	**0.89**
10 dB	PESQ	3.07	3.21	3.19	3.15	**3.26**
	SDR	13.64	16.67	16.22	14.71	**17.34**
	ESTOI	0.90	**0.93**	0.92	0.90	**0.93**

Table 3. Average results for Cafe noise

Input SNR	Eval.	STSA	MNMF -IS-MU	NTF -KL-MU	NTF -IS-MU	VMNMF
-5 dB	PESQ	1.68	1.79	1.73	1.72	**1.87**
	SDR	-0.32	2.66	2.78	1.41	**3.80**
	ESTOI	0.38	0.44	0.44	0.41	**0.48**
0 dB	PESQ	2.01	2.08	2.05	2.04	**2.23**
	SDR	4.70	6.91	6.90	6.01	**8.52**
	ESTOI	0.55	0.59	0.59	0.57	**0.64**
5 dB	PESQ	2.35	2.40	2.39	2.37	**2.58**
	SDR	9.15	10.48	10.30	9.57	**12.38**
	ESTOI	0.71	0.73	0.73	0.71	**0.77**
10 dB	PESQ	2.67	2.71	2.71	2.70	**2.88**
	SDR	13.00	13.51	12.94	12.19	**15.48**
	ESTOI	0.83	0.84	0.83	0.81	**0.86**

Table 4. Average results for Pedestrian noise

Input SNR	Eval.	STSA	MNMF -IS-MU	NTF -KL-MU	NTF -IS-MU	VMNMF
-5 dB	PESQ	1.72	1.89	1.83	1.86	**1.97**
	SDR	-0.49	3.17	2.48	1.93	**4.42**
	ESTOI	0.39	0.47	0.45	0.44	**0.50**
0 dB	PESQ	2.07	2.19	2.16	2.17	**2.33**
	SDR	4.50	7.33	6.80	6.26	**9.05**
	ESTOI	0.56	0.62	0.60	0.59	**0.66**
5 dB	PESQ	2.41	2.51	2.49	2.48	**2.66**
	SDR	8.96	10.92	10.46	9.89	**12.86**
	ESTOI	0.71	0.75	0.74	0.72	**0.79**
10 dB	PESQ	2.72	2.81	2.79	2.77	**2.95**
	SDR	12.90	14.03	13.36	12.57	**15.97**
	ESTOI	0.83	0.85	0.84	0.82	**0.87**

Table 5. Average results for Street noise

Input SNR	Eval.	STSA	MNMF -IS-MU	NTF -KL-MU	NTF -IS-MU	VMNMF
-5 dB	PESQ	1.90	2.06	2.00	1.96	**2.16**
	SDR	0.51	5.13	4.38	4.72	**6.52**
	ESTOI	0.46	0.54	0.51	0.51	**0.57**
0 dB	PESQ	2.25	2.38	2.35	2.29	**2.50**
	SDR	5.11	8.88	8.38	8.27	**10.42**
	ESTOI	0.62	0.68	0.66	0.65	**0.71**
5 dB	PESQ	2.58	2.69	2.67	2.60	**2.81**
	SDR	9.22	12.14	11.86	10.95	**13.49**
	ESTOI	0.76	0.80	0.79	0.77	**0.82**
10 dB	PESQ	2.87	2.98	2.97	2.90	**3.08**
	SDR	12.86	14.87	14.73	12.83	**15.99**
	ESTOI	0.86	0.88	0.88	0.86	**0.90**

of 410 utterances of the 6-channel noisy speech signals. The latter signals were generated by scaling and adding the noise to the filtered clean speech signal to obtain input SNRs of -5, 0, 5, and 10 dB. Specifically, the clean speech signals were filtered by the *time-varying* impulse responses (IR) between the speaker and microphones, where the IR is estimated from the real recorded signals (see [22] for more details about the database).

5.2. Methodology

For the STFT analysis, we used a Hanning window of 1024 samples with 50% overlap. We considered $M = 80$ basis vectors for the clean speech and noises and $L_b = 32$ buffer size. We used the diagonal loading factor of $\lambda = 0.01$ and temporal smoothing factor of $\tau_C = 0.9$. We fixed the shape parameters of the gamma distributions to $\alpha_a = 1$ and $\alpha_h = 10$ in the enhancement stage for simplicity, i.e., to avoid additional iterations as given by (12). Regarding the preprocessor, we implemented the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [23], where the noise PSD was estimated based on [24].

To evaluate the enhancement performance of the proposed VMNMF method, we implemented several benchmark algorithms: i) the MMSE-STSA estimator [23] applied to each channel, ii) the MU-based MNMF method with the IS-divergence (MNMF-IS-MU) [8], iii) the MU-based non-negative tensor factorization (NTF) method with the KL and IS divergences (NTF-KL-MU and NTF-IS-MU) [25]. These benchmarks were essentially used for computing the noise correlation matrix. Note that the MNMF methods (i.e., MNMF-IS-MU and VMNMF) employ the frequency-dependent mixing elements, whereas the NTF methods (i.e., NTF-KL-MU and NTF-IS-MU) use the frequency-independent mixing elements. The MNMF-IS-MU and NTF-IS-MU methods were applied to the power spectra, whereas the NTF-KL-MU method was applied to the magnitude spectra. Basic settings, such as the STFT analysis, the number of basis vectors, the buffer size, and the reconstruction method, were kept identical, except that we used $\lambda = 0.1$ for the benchmarks since it provided better enhancement performance.

We considered the perceptual evaluation of speech quality (PESQ) [26], source-to-distortion ratio (SDR) [27] and extended short-time objective intelligibility (ESTOI) [28] as the objective measures of performance. For all the measures, a higher value indicates a better result.

5.3. Results

The average results over all test utterances are shown in Tables 2 to 5. We can observe that the proposed VMNMF method provided better enhancement performance than the benchmarks for all types of noises and input SNRs. Comparing between the MNMF-IS-MU and NTF-IS-MU methods, the former method provided better results in general. This indicates that the MNMF model can better handle the convolutive effects specified by the ATFs. We also found that the computational cost of the VMNMF method was comparable to that of the efficient MU-based MNMF, NTF-KL and NTF-IS methods. Moreover, the results of an independent series of experiments (not reported due to space limitations), show that among MVDR, post-processed MVDR, and the GM-based reconstruction methods, the latter generally provided better enhancement performance but for a few exceptions.

6. CONCLUSION

We introduced a supervised multi-channel speech enhancement algorithm based on a Bayesian MNMF model. We considered the PGM of MNMF, specified by Poisson-distributed latent variables and gamma-distributed priors. In the training stage, the MNMF parameters of sources were estimated using the VBEM algorithm. In the enhancement stage, the clean speech signal was estimated via the MNMF-based MVDR beamformer. Further improvement of performance was achieved by combining the MNMF-based beamforming technique with a classical unsupervised single-channel enhancement method. Experiments showed that the proposed method provided better enhancement performance than the selected benchmarks.

7. REFERENCES

[1] O. L. Frost, "An algorihtm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926-935, Aug. 1972.

[2] M. Souden, J. Benesty and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925-4935, Sep. 2010.

[3] A. Krueger, E. Warsitz and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 206-219, Jan. 2011.

[4] S. Markovich, S. Gannot and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 6, pp. 1071-1086, June 2009.

[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] 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.

[7] H. Chung, R. Badeau, E. Plourde and B. Champagne, "Training and compensation of class-conditioned NMF bases for speech enhancement," *Neurocomputing*, vol. 284, pp. 107-118, Apr. 2018.

[8] 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.

[9] H. Sawada, H. Kameoka, S. Araki and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 971-982, May 2013.

[10] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 3, pp. 727-739, Mar. 2014.

[11] 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.

[12] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, no. 4, Article ID 785152, pp. 1-17, 2009.

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

[14] X. Mestre and M. A. Launas, "On diagonal loading for minimum variance beamformers," in *Proc. IEEE Int. Symposium on Signal Process. and Information Technology (ISSPIT)*, pp. 459-462, Dec. 2003.

[15] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 1, pp. 223-233, Jan. 2012.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Advances in Neural Information Process. Systems (NIPS)*, pp. 556-562, 2001.

[17] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proc. Int. Joint Conf. Neural Networks*, pp. 486-491, Oct. 2008.

[18] C. Blandin, A. Ozerov and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950-1960, Aug. 2012.

[19] B. Loesch and B. Yang, "Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions," in *Proc. Int. Conf. Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 41-48, Sep. 2010.

[20] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP J. Advances in Signal Process.*, no. 1, p. 61, Dec. 2015.

[21] 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, and Language Process.*, vol. 23, no. 7, pp. 1233-1242, July 2015.

[22] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535-557, Dec. 2016.

[23] 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.

[24] 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.

[25] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *Proc. Int. Symposium on Computer Music Modeling and Retrieval*, pp. 102-115, June 2010.

[26] 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.

[27] 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.

[28] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 2009-2022, Nov. 2016.