Noisy cGMM: Complex Gaussian Mixture Model with Non-Sparse Noise Model for Joint Source Separation and Denoising

Nobutaka Ito, Christopher Schymura, Shoko Araki, and Tomohiro Nakatani
NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
Email: {ito.nobutaka, araki.shoko, nakatani.tomohiro}@lab.ntt.co.jp

***Abstract*—Here we introduce a *noisy cGMM*, a probabilistic model for noisy, mixed signals observed by a microphone array for joint source separation and denoising. In a conventional time-varying complex Gaussian mixture model (cGMM), the observed signals are assumed to be composed of sparse target signals only, where the sparseness refers to the property of having significant power at only a few time-frequency points. However, this assumption becomes inaccurate in the presence of non-sparse signals such as background noise, which renders speech enhancement based on the cGMM less effective. In contrast, the proposed noisy cGMM is based on the assumption that the observed signals consist of not only sparse target signals but also non-sparse background noise. This enables the noisy cGMM to model the observed signals accurately even in the presence of non-sparse background noise, which leads to effective speech enhancement. We also propose a joint diagonalization-based algorithm for estimating the model parameters of the noisy cGMM, which is significantly faster than the standard EM algorithm without any performance degradation. Indeed, the joint diagonalization bypasses the need for matrix inversion, matrix multiplication, and determinant computation at each time-frequency point, which are needed in the EM algorithm. In an experiment, the noisy cGMM outperformed the cGMM in joint source separation and denoising.**

I. Introduction

In this paper, we address the problem of joint source separation and denoising, where the applications include meeting speech recognition in noisy environments.

A promising approach is that exploiting sparseness of source signals, *i.e.,* the property of having significant power at only a few time-frequency points. For example, it is well-known that speech signals are quite sparse. Among all, a time-varying *complex Gaussian mixture model (cGMM)* [1], [2] is known to realize highly effective speech enhancement. The cGMM models the spatial characteristics of each source signal by a full-rank matrix called a *spatial covariance matrix* as in [3]. This enables the cGMM to better model reverberant source signals or diffuse signals than the conventional rank-one modeling using a steering vector. The effectiveness of the cGMM has been demonstrated for source separation [1], [2] and denoising [4]. Especially, it played a central role in the best-performing system [4] in the CHiME-3 challenge [5].

The cGMM relies on the assumption that there are only sparse target signals in the environment. Based on this, the observed signals at each time-frequency point are assumed to be composed of a single target signal [6]. In the cGMM, each Gaussian component models a target signal. However, the above assumption may not be true in the real world involving background noise, because the background noise is often non-sparse. This renders speech enhancement based on the cGMM less effective in the presence of non-sparse background noise.

To overcome this drawback, here we introduce a *noisy cGMM*, a probabilistic model for noisy, mixed observed signals. Unlike the conventional cGMM, the noisy cGMM is based on the assumption that there is not only sparse, disjoint target signals but also non-sparse background noise in the environment. Based on this, the observed signals at each time-frequency point are assumed to be composed of one target signal plus the background noise. In the noisy cGMM, each Gaussian component models one target signal plus the background noise. Specifically, the covariance matrix of each Gaussian component is a convex combination of the spatial covariance matrices of a target signal and the background noise. Consequently, the noisy cGMM can model the observed signals accurately thereby enabling effective speech enhancement even in the presence of non-sparse background noise.

We also present two alternative algorithms for estimating the model parameters of the noisy cGMM. One based on the standard EM algorithm requires matrix inversion, matrix multiplication, and determinant computation at each time-frequency point. However, this may be prohibitive in applications with restricted computational resources (*e.g.*, hearing aids) or to a large dataset (*e.g.*, the CHiME-3 dataset [5]). To resolve this issue, we also propose a joint diagonalization-based acceleration of the EM algorithm, which is significantly faster than the standard EM algorithm without any performance degradation. Indeed, the joint diagonalization bypasses the need for matrix inversion, matrix multiplication, and determinant computation at each time-frequency point, which are needed in the EM algorithm.

The rest of this paper is organized as follows. Section II formulates the problem. Section III introduces the noisy cGMM. Section IV describes the joint diagonalization-based fast algorithm. Section V describes an experiment, and finally, Section VI concludes this paper.

II. Problem Formulation

Suppose $N\,(\geq 1)$ target signals are observed by an array of $M\,(\geq 2)$ microphones in the presence of background noise.

ISBN 978-90-827970-1-5 © EURASIP 2018

Here, the target signals are assumed to be sparse and disjoint, while the background noise is assumed to be non-sparse. We denote the signal observed by the mth microphone by $y_m(t,f)$ in the short-time Fourier transform (STFT) domain. Here, we denote the microphone index by $m \in \{1,\ldots,M\}$, the frame index by $t \in \{1,\ldots,T\}$, and the frequency-bin index by $f \in \{1,\ldots,F\}$. The signals observed by all microphones are collected in a vector $\mathbf{y}(t,f) = \big(y_1(t,f) \quad \ldots \quad y_M(t,f)\big)^T$, which we call an *observation vector*.

The observation vector $\mathbf{y}(t,f) \in \mathbb{C}^M$ is composed of N components $\mathbf{x}_1(t,f),\ldots,\mathbf{x}_N(t,f) \in \mathbb{C}^M$ corresponding to the N target signals plus a component $\mathbf{v}(t,f)$ corresponding to the background noise:

$$\mathbf{y}(t,f) = \sum_{\nu=1}^{N} \mathbf{x}_\nu(t,f) + \mathbf{v}(t,f), \tag{1}$$

where $\nu \in \{1,\ldots,N\}$. We refer to $\mathbf{x}_1(t,f),\ldots,\mathbf{x}_N(t,f)$ and $\mathbf{v}(t,f)$ as *source images*. Specifically, $\mathbf{x}_1(t,f),\ldots,\mathbf{x}_N(t,f)$ corresponding to the target signals are called *target images*, and $\mathbf{v}(t,f)$ corresponding to the background noise is called a *noise image*.

As already discussed, we assume that the target signals are sparse and disjoint but the background noise is non-sparse. Based on these assumptions, we approximate (1) by

$$\mathbf{y}(t,f) \simeq \underbrace{\mathbf{x}_{n(t,f)}(t,f)}_{\text{target signal}} + \underbrace{\mathbf{v}(t,f)}_{\text{noise}}, \tag{2}$$

where $n(t,f) \in \{1,\ldots,N\}$ denotes the index of the target signal present at (t,f). Equation (2) implies that the observed signals are composed of a single target signal plus the background noise at each time-frequency point.

The problem of joint source separation and denoising we deal with is one of estimating each target signal given the observation vector $\mathbf{y}(t,f)$.

III. Noisy cGMM for Joint Source Separation and Denoising

A. Probabilistic Modeling with Noisy cGMM

As in [3], we model the probability distribution of each source image by a time-varying complex Gaussian distribution:

$$p(\mathbf{x}_\nu(t,f)) = \mathcal{N}(\mathbf{x}_\nu(t,f); \mathbf{0}, \mathbf{S}_\nu(t,f)), \tag{3}$$
$$p(\mathbf{v}(t,f)) = \mathcal{N}(\mathbf{v}(t,f); \mathbf{0}, \mathbf{S}_0(t,f)), \tag{4}$$

where $\mathcal{N}(\boldsymbol{\eta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Each covariance matrix $\mathbf{S}_i(t,f)$ $(i = 0,\ldots,N)$ is parameterized by a *spatial covariance matrix* $\mathbf{R}_i(f)$ and a *power parameter* $v_i(t,f)$, where the former models the spatial characteristics of each source signal, and the latter models its power spectrum [3]:

$$\mathbf{S}_i(t,f) = \underbrace{v_i(t,f)}_{\text{power spectrum}} \times \underbrace{\mathbf{R}_i(f).}_{\text{spatial characteristics}} \tag{5}$$

We assume that $v_i(t,f)$ is time-variant and positive and $\mathbf{R}_i(f)$ is time-invariant and Hermitian positive definite. Here, we used the index i instead of ν to highlight that it can take zero.

From the reproduction property of the complex Gaussian distribution along with (2)–(5), we have the following mixture model, which we call a *noisy cGMM*:

$$\begin{aligned} &p(\mathbf{y}(t,f)) \qquad (6)\\ &= \sum_{\nu=1}^{N} \alpha_\nu(f) \mathcal{N}(\mathbf{y}(t,f); \mathbf{0}, \underbrace{v_\nu(t,f)\mathbf{R}_\nu(f)}_{\text{target signal}} + \underbrace{v_0(t,f)\mathbf{R}_0(f)}_{\text{noise}}). \end{aligned}$$

The noisy cGMM (6) is said to be 'noisy' because it has the noise term $v_0(t,f)\mathbf{R}_0(f)$ accounting for the non-sparse background noise. The mixture weight $\alpha_\nu(f)$ models the prior probability of the presence of the νth target signal, and satisfies $\sum_{\nu=1}^{N} \alpha_\nu(f) = 1$.

B. EM Algorithm

The model parameters of the noisy cGMM, namely $\alpha_\nu(f)$, $v_i(t,f)$, and $\mathbf{R}_i(f)$, can be estimated in the maximum-likelihood sense based on the EM algorithm, in which the E-step and the M-step are alternated. The target image present at (t,f) (see (2)) denoted by

$$\mathbf{x}(t,f) \triangleq \mathbf{x}_{n(t,f)}(t,f) \tag{7}$$

and its index $n(t,f)$ are considered to be hidden variables.

In the E-step, the posterior distribution $p(\mathbf{x}(t,f), n(t,f) = \nu \mid \mathbf{y}(t,f))$ of the hidden variables is updated based on the current parameter estimates. This posterior distribution is decomposed as

$$\begin{aligned} &p(\mathbf{x}(t,f), n(t,f) = \nu \mid \mathbf{y}(t,f)) \qquad (8)\\ &= \underbrace{P(n(t,f) = \nu \mid \mathbf{y}(t,f))}_{\gamma_\nu(t,f)} p(\mathbf{x}(t,f) \mid n(t,f) = \nu, \mathbf{y}(t,f)), \end{aligned}$$

where $\gamma_\nu(t,f)$ is the posterior probability of the presence of the νth target signal, and can be regarded as a mask. Since the second factor of (8) turns out to be a complex Gaussian distribution $\mathcal{N}(\mathbf{x}(t,f); \boldsymbol{\mu}_\nu(t,f), \boldsymbol{\Sigma}_\nu(t,f))$, the E-step amounts to updating its mean $\boldsymbol{\mu}_\nu(t,f)$ and covariance matrix $\boldsymbol{\Sigma}_\nu(t,f)$ along with the masks $\gamma_\nu(t,f)$.

The M-step updates the parameter estimates so as to increase an auxiliary Q function, which is defined using the posterior distribution updated in the E-step.

This EM algorithm is theoretically guaranteed to increase the likelihood function monotonically. The algorithm is shown in Algorithm 1, where $\nu \in \{1,\ldots,N\}$, $i \in \{0,\ldots,N\}$, and $\gamma_0(t,f) \triangleq 1$.

C. Estimation of Target Signals

Once the parameters have been estimated, the target signals can be estimated in various ways. For example, this can be done using minimum variance distortionless response (MVDR) beamformers, where the steering vectors are estimated based on the generalized eigenvalue problem of the estimated spatial covariance matrices $\mathbf{R}_\nu(f)$ and $\mathbf{R}_0(f)$ [7],

Algorithm 1. EM algorithm for the noisy cGMM.

```
Initialize the parameters $\alpha_\nu(f)$, $v_i(t,f)$, and $\mathbf{R}_i(f)$.
repeat
    % E-step
    $\gamma_\nu(t,f) \leftarrow \dfrac{\alpha_\nu(f)\mathcal{N}(\mathbf{y}(t,f);\mathbf{0},v_\nu(t,f)\mathbf{R}_\nu(f)+v_0(t,f)\mathbf{R}_0(f))}{\sum_{\nu'=1}^{N}\alpha_{\nu'}(f)\mathcal{N}(\mathbf{y}(t,f);\mathbf{0},v_{\nu'}(t,f)\mathbf{R}_{\nu'}(f)+v_0(t,f)\mathbf{R}_0(f))},\ {}^\forall\nu,t,f.$
    $\boldsymbol{\mu}_\nu(t,f) \leftarrow v_\nu(t,f)\mathbf{R}_\nu(f)(v_\nu(t,f)\mathbf{R}_\nu(f)+v_0(t,f)\mathbf{R}_0(f))^{-1}\mathbf{y}(t,f),\ {}^\forall\nu,t,f.$
    $\boldsymbol{\Sigma}_\nu(t,f) \leftarrow v_\nu(t,f)\mathbf{R}_\nu(f)(v_\nu(t,f)\mathbf{R}_\nu(f)+v_0(t,f)\mathbf{R}_0(f))^{-1}(v_0(t,f)\mathbf{R}_0(f)),\ {}^\forall\nu,t,f.$
    $\boldsymbol{\Phi}_\nu(t,f) \leftarrow \boldsymbol{\Sigma}_\nu(t,f)+\boldsymbol{\mu}_\nu(t,f)\boldsymbol{\mu}_\nu(t,f)^H,\ {}^\forall\nu,t,f.$
    $\boldsymbol{\Phi}_0(t,f) \leftarrow \sum_{\nu=1}^{N}\gamma_\nu(t,f)\Big[\boldsymbol{\Sigma}_\nu(t,f)+(\mathbf{y}(t,f)-\boldsymbol{\mu}_\nu(t,f))(\mathbf{y}(t,f)-\boldsymbol{\mu}_\nu(t,f))^H\Big],\ {}^\forall t,f.$
    % M-step
    $\alpha_\nu(f) \leftarrow \dfrac{1}{T}\sum_{t=1}^{T}\gamma_\nu(t,f),\ {}^\forall\nu,f.$
    $v_i(t,f) \leftarrow \dfrac{1}{M}\mathrm{tr}(\mathbf{R}_i(f)^{-1}\boldsymbol{\Phi}_i(t,f)),\ {}^\forall i,t,f.$
    $\mathbf{R}_i(f) \leftarrow \dfrac{1}{\sum_{t=1}^{T}\gamma_i(t,f)}\sum_{t=1}^{T}\gamma_i(t,f)\dfrac{1}{v_i(t,f)}\boldsymbol{\Phi}_i(t,f),\ {}^\forall i,f.$
until a predetermined end condition is met.
```

TABLE I
COMPARISON OF THE NUMBER OF MATRIX OPERATIONS OF COMPLEXITY $O(M^3)$ PER ITERATION.

	EM algorithm	JEM algorithm
$\mathbf{A}^{-1}$	$(TN+N+1)F$	FN
$\mathbf{AB}$	$2TFN$	$4FN$
$\det\mathbf{A}$	TFN	0
$\mathbf{Ap}=\lambda\mathbf{Bp}$	0	FN
Total	$(4TN+N+1)F$	$6FN$

[8]. To further suppress the residual noise at the output of the MVDR beamformers, post-filtering can be performed using the estimated masks $\gamma_\nu(t,f)$.

IV. JOINT DIAGONALIZATION BASED ACCELERATED EM ALGORITHM FOR NOISY cGMM

A. Motivation

A drawback of Algorithm 1 is expensive computation. Especially, the E-step requires matrix inversion, matrix multiplication, and determinant computation, which are of complexity $O(M^3)$, at each time-frequency point at each iteration. Table I shows the number of matrix operations of complexity $O(M^3)$ per iteration in Algorithm 1 (see the column labeled 'EM algorithm') as a rough estimate of the computational complexity.

B. Approach: Joint Diagonalization

In this section, we propose a joint diagonalization-based acceleration of the EM algorithm in Algorithm 1. As in [9], [10], we exploit the fact that, for diagonal matrices, matrix inversion, matrix multiplication, and determinant computation are reduced to mere scalar operations of the diagonal entries, which are of complexity $O(M)$ instead of $O(M^3)$. This implies that, if the covariance matrices $\mathbf{S}_\nu(t,f)=v_\nu(t,f)\mathbf{R}_\nu(f)$ and $\mathbf{S}_0(t,f)=v_0(t,f)\mathbf{R}_0(f)$ were both diagonal, matrix inversion, matrix multiplication, and determinant computation in Algorithm 1 could be performed in $O(M)$. In practice, however, $\mathbf{S}_\nu(t,f)$ and $\mathbf{S}_0(t,f)$ are not diagonal in most cases. This is because different entries of a source image, which correspond to different microphones, are usually highly correlated.

This motivates us to consider the joint diagonalization of $\mathbf{R}_\nu(f)$ and $\mathbf{R}_0(f)$ based on the generalized eigenvalue problem. Let $\lambda_{\nu1}(f),\cdots,\lambda_{\nu M}(f)$ be the generalized eigenvalues, and $\mathbf{p}_{\nu1}(f),\cdots,\mathbf{p}_{\nu M}(f)$ be generalized eigenvectors corresponding to $\lambda_{\nu1}(f),\cdots,\lambda_{\nu M}(f)$, respectively, so that

$$\mathbf{R}_\nu(f)\mathbf{p}_{\nu m}(f)=\lambda_{\nu m}(f)\mathbf{R}_0(f)\mathbf{p}_{\nu m}(f). \tag{9}$$

The generalized eigenvectors $\mathbf{p}_{\nu1}(f),\cdots,\mathbf{p}_{\nu M}(f)$ can be taken so that they satisfy the following orthonormality:

$$\mathbf{p}_{\nu l}(f)^H\mathbf{R}_0(f)\mathbf{p}_{\nu m}(f)=\delta_{lm}, \tag{10}$$

where δ_{lm} denotes the Kronecker delta. Rewriting these equations in matrix form, we have

$$\begin{cases}\mathbf{P}_\nu(f)^H\mathbf{R}_\nu(f)\mathbf{P}_\nu(f)=\boldsymbol{\Lambda}_\nu(f),\\ \mathbf{P}_\nu(f)^H\mathbf{R}_0(f)\mathbf{P}_\nu(f)=\mathbf{I},\end{cases} \tag{11}$$

Algorithm 2. JEM algorithm for the noisy cGMM.

```
Initialize the parameters $\alpha_\nu(f)$, $v_i(t,f)$, and $\mathbf{R}_i(f)$.
repeat
    % J-step
    For each $f$ and $\nu$, compute a non-singular matrix $\mathbf{P}_\nu(f)$ and a diagonal matrix $\mathbf{\Lambda}_\nu(f)$ satisfying (11) by solving the generalized eigenvalue problem of $\mathbf{R}_\nu(f)$ and $\mathbf{R}_0(f)$.
    % E-step
    $\mathbf{y}'_\nu(t,f) \leftarrow \mathbf{P}_\nu(f)^H \mathbf{y}(t,f), \quad {}^\forall\nu,t,f.$
    $\gamma_\nu(t,f) \leftarrow \dfrac{\alpha_\nu(f)\mathcal{N}(\mathbf{y}'_\nu(t,f);\mathbf{0},v_\nu(t,f)\mathbf{\Lambda}_\nu(f)+v_0(t,f)\mathbf{I})}{\sum_{\nu'=1}^{N}\alpha_{\nu'}(f)\mathcal{N}(\mathbf{y}'_{\nu'}(t,f);\mathbf{0},v_{\nu'}(t,f)\mathbf{\Lambda}_{\nu'}(f)+v_0(t,f)\mathbf{I})}, \quad {}^\forall\nu,t,f.$
    $\boldsymbol{\mu}'_\nu(t,f) \leftarrow v_\nu(t,f)\mathbf{\Lambda}_\nu(f)(v_\nu(t,f)\mathbf{\Lambda}_\nu(f)+v_0(t,f)\mathbf{I})^{-1}\mathbf{y}'_\nu(t,f), \quad {}^\forall\nu,t,f.$
    $\mathbf{\Sigma}'_\nu(t,f) \leftarrow v_0(t,f)v_\nu(t,f)\mathbf{\Lambda}_\nu(f)(v_\nu(t,f)\mathbf{\Lambda}_\nu(f)+v_0(t,f)\mathbf{I})^{-1}, \quad {}^\forall\nu,t,f.$
    % M-step
    $\alpha_\nu(f) \leftarrow \dfrac{1}{T}\sum_{t=1}^{T}\gamma_\nu(t,f), \quad {}^\forall\nu,f.$
    $v_\nu(t,f) \leftarrow \dfrac{1}{M}\mathrm{tr}\big[\mathbf{\Lambda}_\nu(f)^{-1}(\mathbf{\Sigma}'_\nu(t,f)+\boldsymbol{\mu}'_\nu(t,f)\boldsymbol{\mu}'_\nu(t,f)^H)\big], \quad {}^\forall\nu,t,f.$
    $v_0(t,f) \leftarrow \dfrac{1}{M}\mathrm{tr}\Bigg\{\sum_{\nu=1}^{N}\gamma_\nu(t,f)\big[\mathbf{\Sigma}'_\nu(t,f)+(\mathbf{y}'_\nu(t,f)-\boldsymbol{\mu}'_\nu(t,f))(\mathbf{y}'_\nu(t,f)-\boldsymbol{\mu}'_\nu(t,f))^H\big]\Bigg\}, \quad {}^\forall t,f.$
    $\mathbf{R}_\nu(f) \leftarrow (\mathbf{P}_\nu(f)^{-1})^H\left[\dfrac{1}{\sum_{t=1}^{T}\gamma_\nu(t,f)}\sum_{t=1}^{T}\gamma_\nu(t,f)\dfrac{1}{v_\nu(t,f)}(\mathbf{\Sigma}'_\nu(t,f)+\boldsymbol{\mu}'_\nu(t,f)\boldsymbol{\mu}'_\nu(t,f)^H)\right]\mathbf{P}_\nu(f)^{-1}, \quad {}^\forall\nu,f.$
    $\mathbf{R}_0(f) \leftarrow \sum_{\nu=1}^{N}(\mathbf{P}_\nu(f)^{-1})^H\left[\dfrac{1}{T}\sum_{t=1}^{T}\dfrac{1}{v_0(t,f)}\gamma_\nu(t,f)(\mathbf{\Sigma}'_\nu(t,f)+(\mathbf{y}'_\nu(t,f)-\boldsymbol{\mu}'_\nu(t,f))(\mathbf{y}'_\nu(t,f)-\boldsymbol{\mu}'_\nu(t,f))^H)\right]\mathbf{P}_\nu(f)^{-1},$
        ${}^\forall f.$
until a predetermined end condition is met.
```

where $\mathbf{P}_\nu(f)$ denotes the matrix composed of $\mathbf{p}_{\nu 1}(f),\cdots,\mathbf{p}_{\nu M}(f)$, and $\mathbf{\Lambda}_\nu(f)$ the diagonal matrix composed of $\lambda_{\nu 1}(f),\cdots,\lambda_{\nu M}(f)$. Equation (11) implies that $\mathbf{P}_\nu(f)$ diagonalizes $\mathbf{R}_\nu(f)$ and $\mathbf{R}_0(f)$ jointly.

C. Accelerated Algorithm: JEM Algorithm

Algorithm 2 shows the joint diagonalization based accelerated EM algorithm for the noisy cGMM, which we call a *JEM (joint diagonalization-expectation-maximization) algorithm*. As in Algorithm 1, $\nu \in \{1,\ldots,N\}$ and $i \in \{0,\ldots,N\}$. The JEM algorithm consists of a J-step, an E-step, and an M-step.

The E-step of Algorithm 2 is obtained by solving (11) for $\mathbf{R}_\nu(f)$ and $\mathbf{R}_0(f)$, plugging them in the E-step of Algorithm 1, and defining $\boldsymbol{\mu}'_\nu(t,f)$, $\mathbf{\Sigma}'_\nu(t,f)$, and $\mathbf{y}'_\nu(t,f)$ by

$$\boldsymbol{\mu}'_\nu(t,f) \triangleq \mathbf{P}_\nu(f)^H\boldsymbol{\mu}_\nu(t,f), \tag{12}$$

$$\mathbf{\Sigma}'_\nu(t,f) \triangleq \mathbf{P}_\nu(f)^H\mathbf{\Sigma}_\nu(t,f)\mathbf{P}_\nu(f), \tag{13}$$

$$\mathbf{y}'_\nu(t,f) \triangleq \mathbf{P}_\nu(f)^H\mathbf{y}(t,f). \tag{14}$$

The M-step of Algorithm 2 is obtained by solving (12), (13), and (14) for $\boldsymbol{\mu}_\nu(t,f)$, $\mathbf{\Sigma}_\nu(t,f)$, and $\mathbf{y}(t,f)$, and plugging them into the M-step of Algorithm 1.

D. Advantage of JEM Algorithm

The joint diagonalization bypasses the need for time-frequency-wise matrix operations of complexity $O(M^3)$. As a result, Algorithm 2 requires a significantly reduced number of matrix operations of complexity $O(M^3)$ compared to Algorithm 1 (see Table I). In Table I, we do not count matrix inversions, matrix multiplications, and determinant computations of diagonal matrices, which are of complexity $O(M)$ instead of $O(M^3)$. Although the JEM algorithm requires additional generalized eigenvalue decompositions (complexity $O(M^3)$), they are not required frame-wise. Therefore, the total number of matrix operations of complexity $O(M^3)$ is significantly less than that in the EM algorithm. Indeed,

$$\frac{(4TN+N+1)F}{6FN} = \frac{4T+1}{6} + \frac{1}{6N} \gg 1 \tag{15}$$

because $T \gg 1$.

V. Experiment: Joint Source Separation and Denoising

We conducted an experiment of joint source separation and denoising to compare the performance of the proposed noisy

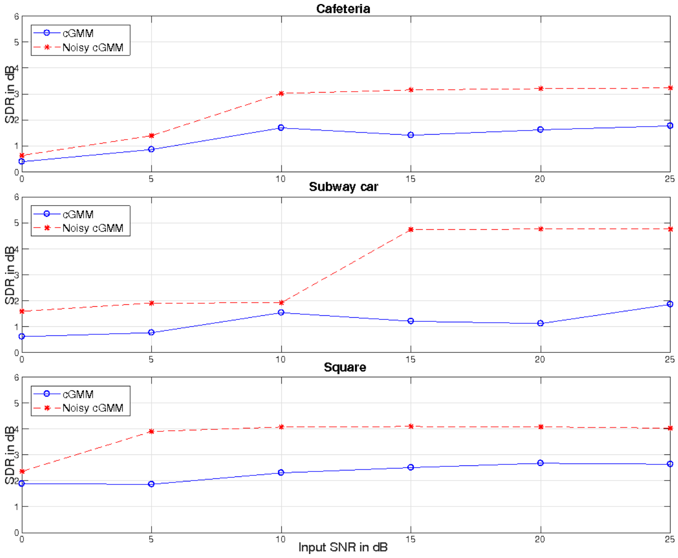

Fig. 1. Performance in terms of the signal-to-distortion ratio (SDR) for three noise environments: ”Cafeteria”, ”Subway car”, and ”Square”.

cGMM and the conventional cGMM.

The observed signals were generated by remixing the target images and the noise image in the ”Dev” set of the task ”source separation in the presence of real-world background noise” proposed in SiSEC2010 [11]. The target signals were $N = 3$ speech signals, which were observed by a uniform linear array with $M = 4$ elements. See [11] for the details of the database.

For the noisy cGMM, the JEM algorithm (Algorithm 2) was employed. Note that Algorithm 2 is equivalent to Algorithm 1 and yields virtually the identical performance, though computationally much less expensive. The conventional cGMM had three clusters: four clusters with an additional noise cluster did not work in our experiment. In both methods, $\mathbf{R}_\nu(f)$ were initialized based on a conventional clustering method [12]. In the noisy cGMM, $\mathbf{R}_0(f)$ was initialized based on a spherically isotropic noise model [13]. For both methods, the target signals were estimated by MVDR beamforming followed by post-filtering as described in Section III-C. For the conventional cGMM, the spherically isotropic noise model was employed for $\mathbf{R}_0(f)$ in steering vector estimation. For both methods, permutation alignment was performed based on the conventional clustering method [12].

Figure 1 shows the performance in terms of signal-to-distortion ratio (SDR) [14] as a function of the input SNR. The input SNR is defined as the power ratio between a target signal and the background noise, averaged for all target signals. The SDR was evaluated between the estimated target signal and the true target signal at the reference microphone. We see that the noisy cGMM consistently outperformed the conventional cGMM.

VI. Conclusions

In this paper, we proposed the noisy cGMM for joint source separation and denoising. The future work includes the application of the noisy cGMM to meeting speech recognition in a real-world noisy environment.

References

[1] R. Sakanashi, S. Miyabe, T. Yamada, and S. Makino, “Comparison of superimposition and sparse models in blind source separation by multichannel Wiener filter,” in *Proc. APSIPA*, Dec. 2012.

[2] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, “Relaxed disjointness based clustering for joint blind source separation and dereverberation,” in *Proc. IWAENC*, Sept. 2014, pp. 268–272.

[3] N. Q. K. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[4] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, “The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” in *Proc. ASRU*, Dec. 2015, pp. 436–443.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *Proc. ASRU*, Dec. 2015, pp. 504–511.

[6] Ö. Yılmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. SP*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[7] E. Warsitz and R. Heab-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” *IEEE Trans. ASLP*, vol. 15, no. 5, pp. 1529–1439, June 2007.

[8] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, “Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments,” in *Proc. ICASSP*, Mar. 2017, pp. 681–685.

[9] N. Ito, S. Araki, and T. Nakatani, “FastFCA: A joint diagonalization based fast algorithm for audio source separation using a full-rank spatial covariance model,” *arXiv preprint*, May 2018, arXiv: 1805.06572.

[10] ——, “FastFCA: Joint diagonalization based acceleration of audio source separation using a full-rank spatial covariance model,” in *Proc. EUSIPCO*, Sept. 2018.

[11] S. Araki, A. Ozerov, B. V. Gowreesunker, H. Sawada, F. J. Theis, G. Nolte, D. Lutter, and N. Q. K. Duong, “The 2010 signal separation evaluation campaign (SiSEC2010): Audio source separation,” in *Proc. LVA/ICA*, Sept. 2010, pp. 114–122.

[12] H. Sawada, S. Araki, and S. Makino, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[13] R. Cook, R. Waterhouse, R. Berendt, S. Edelman, and M. Thompson Jr., “Measurement of correlation coefficients in reverberant sound fields,” *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, Nov. 1955.

[14] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.