A Robust Maximum Likelihood Distortionless Response Beamformer based on a Complex Generalized Gaussian Distribution

Weixin Meng*† Chengshi Zheng*† and Xiaodong Li*†
* Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
† University of Chinese Academy of Sciences, Beijing, China
E-mail: {mengweixin, cszheng, lxd}@mail.ioa.ac.cn

***Abstract*—For speech enhancement applications, this paper derives a robust maximum likelihood distortionless response (MLDR) beamformer by introducing a complex generalized Gaussian distribution to model speech sparse priors, where we refer to as the CGGD-MLDR beamformer. In addition, a steering vector estimation method is integrated into the iteration framework of the proposed CGGD-MLDR beamformer. We show that the proposed beamformer can be regarded as a generalization of the minimum power distortionless response beamformer and its improved variations. For narrowband applications, we also reveal that the proposed beamformer reduces to the minimum dispersion distortionless response beamformer, which has been derived with the ℓ_p-norm minimization. The mechanisms of the proposed beamformer in improving the robustness are clearly pointed out and experimental results show its better performance for microphone array speech enhancement in terms of PESQ and ESTOI improvements.**

I. Introduction

Array beamforming has been widely used to extract the desired signal and suppress both the interferences and the noise for many applications, such as sonar, radar, and speech communication systems. Typically, there are two types of microphone array beamforming algorithms, where one is data-independent fixed beamformers and the other is data-dependent adaptive beamformers. Generally, adaptive beamformers are more powerful in suppressing directional interferences automatically than the fixed beamformers. There are many well-known adaptive beamformers including the minimum power distortions response (MPDR) beamformer [1]–[3], the generalized sidelobe cancellation (GSC) [4]–[6] and the multi-channel Wiener filtering (MWF) [7]–[11]. Both the MPDR beamformer and the GSC are quite sensitive to the steering vector error of the desired speech, while the MWF is sensitive to the estimation accuracy of the second-order statistics of speech and noise. Among these beamformers, the MPDR beamformer is still a hot topic for speech enhancement because of its promising performance.

There are at least two ways to improve the robustness of the MPDR beamformer, where one is to improve the estimation accuracy of the steering vector of the desired speech [12]–[16] and the other is to estimate the noise power spectral density (PSD) matrix to replace the noisy PSD matrix [17]–[22]. For practical applications, these two ways can be combined together to further improve the performance. Whereas, one cannot expect that the steering vector of the desired speech can be estimated accurately and the noise PSD matrix does not contain any desired speech PSD matrix, especially in low signal-to-interference-plus-noise ratio (SINR) conditions. This paper focuses on improving the robustness of the MPDR beamformer by estimating a weighted noisy PSD matrix and the steering vector simultaneously under maximum likelihood criterion.

When assuming that the desired speech in the frequency domain follows a zero-mean complex Gaussian distribution (CGD) with time-varying variances, a maximum likelihood distortionless response (MLDR) beamformer was derived in [23] and it can reduce the word error rates for automatic speech recognition. When considering the signal and the noise are non-Gaussian distributed, a minimum dispersion distortionless response (MDDR) beamformer was derived with the ℓ_p-norm minimization for narrowband applications in [24], [25]. The relationship between the MLDR beamformer and the MDDR beamformer has not been revealed clearly and the mechanism in improving performance needs to be further clarified. Moreover, the best choice of p in MDDR is not so straightforward, which also needs to be studied in a more theoretical way.

This paper derives a robust maximum likelihood distortionless response beamformer by introducing a zero-mean complex generalized Gaussian distribution to model speech sparse priors [26], [27], which is referred to as the CGGD-MLDR beamformer. Meanwhile, we present a steering vector estimation method in the iteration framework of the proposed beamformer. One can see that the proposed beamformer is a generalization of the MPDR beamformer and it can reduce to many existing variations of the MPDR beamformer. After revealing the relationship of the proposed CGGD-MLDR beamformer, this paper shows the mechanism of the CGGD-MLDR in improving the robustness of the MPDR beamformer. Moreover, to implement the proposed beamformer, we design an iterative optimization algorithm to estimate its optimal weight vector and the steering vector alternately. Experimental results show that the proposed CGGD-MLDR beamformer can achieve better performance by properly choosing the shape parameter p.

The remainder of this paper is organized as follows. Section

II presents problem formulation and related work. In section III, we derive the CGGD-MLDR beamformer and study its relationship with the MPDR beamformer and its variations. The mechanism in improving robustness is also presented. In section IV, we study the performance of the CGGD-MLDR beamformer and compare it with the MLDR and the MPDR beamformers. Section V presents some conclusions.

II. Problem Formulation and Related Work

We assume that a desired speech source and some uncorrelated directional noise sources impinge on an arbitrary shape microphone array consisting of M microphones. By applying the short-time Fourier transform (STFT), the microphone array signals can be written into vectors of M in time-frequency bin, denoted by $\mathbf{y}(k,l) = [Y_1(k,l), ..., Y_M(k,l)]^T$, where k indicates the frequency index and l indicates the frame index. We have

$$\begin{aligned}\mathbf{y}(k,l) &= \mathbf{h}(k)S(k,l) + \mathbf{v}(k,l) \\ &= \mathbf{x}(k,l) + \mathbf{v}(k,l),\end{aligned} \tag{1}$$

where $S(k,l)$ denotes the complex spectrum of the desired speech; $\mathbf{h}(k)$ denotes the steering vector of the desired speech; $\mathbf{v}(k,l)$ denotes the noise vector.

The objective of a beamformer is to design a spatial filter $\mathbf{w}(k)$, which can be applied to extract the desired speech:

$$\begin{aligned}\widehat{S}(k,l) &= \mathbf{w}^H(k)\mathbf{y}(k,l) \\ &= \mathbf{w}^H(k)\mathbf{h}(k)S(k,l) + \mathbf{w}^H(k)\mathbf{v}(k,l).\end{aligned} \tag{2}$$

The well-known MPDR beamformer aims to minimize the beamforming output power with the distortionless constraint on the desired direction, which can be given by:

$$\min_{\mathbf{w}(k)} E\left\{\mathbf{w}^H(k)\mathbf{y}(k,l)\right\}^2 \quad s.t. \quad \mathbf{w}^H(k)\mathbf{h}(k) = 1. \tag{3}$$

The close-formed solution of (3) can be written as

$$\mathbf{w}_{\text{MPDR}}(k) = \frac{\mathbf{R}_{\mathbf{yy}}^{-1}(k)\mathbf{h}(k)}{\mathbf{h}^H(k)\mathbf{R}_{\mathbf{yy}}^{-1}(k)\mathbf{h}(k)}, \tag{4}$$

where

$$\begin{aligned}\mathbf{R}_{\mathbf{yy}}(k) &= E\left\{\mathbf{y}(k,l)\mathbf{y}^H(k,l)\right\} \\ &= \lambda_s(k)\boldsymbol{\Upsilon}_{ss}(k) + \lambda_v(k)\boldsymbol{\Upsilon}_{\mathbf{vv}}(k) \\ &= \mathbf{R}_{\mathbf{ss}}(k) + \mathbf{R}_{\mathbf{vv}}(k)\end{aligned} \tag{5}$$

denotes the noisy PSD matrix; $\lambda_s(k)$ denotes the PSD of the desired speech; $\boldsymbol{\Upsilon}_{ss}(k)$ denotes the desired speech correlation matrix; $\lambda_v(k)$ denotes the PSD of the noise signal and $\boldsymbol{\Upsilon}_{\mathbf{vv}}(k)$ denotes the noise correlation matrix; $\mathbf{R}_{\mathbf{ss}}(k)$ and $\mathbf{R}_{\mathbf{vv}}(k)$ denote the desired speech PSD matrix and the noise PSD matrix, respectively. In practice, the noisy PSD matrix needs to be replaced by using its sample covariance matrix, which can be given by:

$$\widehat{\mathbf{R}}_{\mathbf{yy}}(k) = \sum_{l=1}^{\mathcal{L}} \mathbf{y}(k,l)\,\mathbf{y}^H(k,l). \tag{6}$$

Accordingly, the estimated weight vector of the MPDR beamformer can be expressed as

$$\widehat{\mathbf{w}}_{\text{MPDR}}(k) = \frac{\widehat{\mathbf{R}}_{\mathbf{yy}}^{-1}(k)\mathbf{h}(k)}{\mathbf{h}^H(k)\widehat{\mathbf{R}}_{\mathbf{yy}}^{-1}(k)\mathbf{h}(k)}. \tag{7}$$

It is well-known that the MPDR beamformer is sensitive to the estimation error of steering vector $\mathbf{h}(k)$ and the cancellation of the desired speech often occurs. To improve its robustness, the noisy PSD matrix should be replaced by the noise PSD matrix and the steering vector $\mathbf{h}(k)$ should be estimated more accurately. For this purpose, one needs to distinguish noise-only time-frequency bins from noisy bins or the desired speech presence probability (SPP) in each time-frequency bin needs to be estimated before updating the noise PSD matrix. With noise PSD matrix and noisy PSD matrix, we can estimate the steering vector of the desired speech in theory, so that the estimated weight vector of a more robust MPDR beamformer can be obtained finally.

In [23], the MLDR beamformer is derived and its weight vector has the same form as the MPDR beamformer and the only difference is that the noisy PSD matrix in the MPDR beamformer is estimated and replaced by a weighted sample covariance matrix, which can be given by:

$${}_{\text{CGD}}\widehat{\mathbf{R}}_{\mathbf{yy}}(k) = \sum_{l=1}^{\mathcal{L}} \frac{\mathbf{y}(k,l)\,\mathbf{y}^H(k,l)}{\lambda_s(k,l)}, \tag{8}$$

where $\lambda_s(k,l) = E\left\{|S^2(k,l)|\right\}$ indicates the PSD of the desired speech in each time-frequency bin. Note that $|S^2(k,l)|$ is often unknown prior and it needs to be estimated and replaced by its estimated value. That is to say, $\widehat{\lambda}_s(k,l) = |\widehat{S}^2(k,l)|$ should be used for practical applications.

III. Method

A. MLDR with a Complex Generalized Gaussian Distribution

We assume that $S(k,l)$ follows a zero-mean complex generalized Gaussian distribution, and so its probability density function can be expressed as:

$$\rho\left(S(k,l)\right) = \frac{p}{2\pi\gamma\Gamma\left(2/p\right)} e^{-\frac{|S(k,l)|^p}{\gamma^{p/2}}}, \tag{9}$$

where $\gamma > 0$ is the scale parameter, p is the shape parameter of this complex generalized Gaussian distribution, and $\Gamma(\cdot)$ is the Gamma function. Generally, the complex generalized Gaussian can be divided into three groups including super-Gaussian, Gaussian, and sub-Gaussian, respectively. In this letter, the super-Gaussian distribution is considered for speech applications, i.e. $0 < p < 2$. In this case, according to convex analysis, $\rho\left(S(k,l)\right)$ can be represented as a maximization over scaled Gaussians with different variances, which can be expressed as:

$$\rho\left(S(k,l)\right) = \max_{\lambda_s(k,l)>0} \mathbb{N}_{\mathbb{C}}\left(S(k,l); 0, \lambda_s(k,l)\right)\psi\left(\lambda_s(k,l)\right), \tag{10}$$

where $\mathbb{N}_{\mathbb{C}}(S(k,l); 0, \lambda_s(k,l))$ denotes a complex Gaussian distribution with zero-mean and time-varying variance $\lambda_s(k,l)$; $\psi(\cdot)$ denotes a scaling function which is related to

distribution. With this model, the weight vector $\mathbf{w}(k)$ can be optimized by maximizing the following likelihood function:

$$\begin{aligned}&\max_{\mathbf{w}(k)} \prod_{l=1}^{\mathcal{L}} \max_{\lambda_s(k,l)>0} \mathcal{N}_{\mathbb{C}}\left(S\left(k,l\right);0,\lambda_s\left(k,l\right)\right)\psi\left(\lambda_s\left(k,l\right)\right)\\ &s.t. \quad \mathbf{w}^{\mathrm{H}}\left(k\right)\mathbf{h}\left(k\right)=1\end{aligned} \tag{11}$$

which is equivalent to minimize the negative log-likelihood with the weight vector $\mathbf{w}(k)$ and $\lambda_s(k,l)$, which can be expressed as:

$$\begin{aligned}&\min_{\mathbf{w}(k),\lambda_s(k,l)>0} \sum_{l=1}^{\mathcal{L}} \left(\frac{\left|S\left(k,l\right)\right|^2}{\lambda_s\left(k,l\right)}+\log\left(\pi\lambda_s\left(k,l\right)\right)-\log\psi\left(\lambda_s\left(k,l\right)\right)\right)\\ &s.t. \quad \mathbf{w}^{\mathrm{H}}\left(k\right)\mathbf{h}\left(k\right)=1\end{aligned} \tag{12}$$

It is worth noting that the optimization problem depends on $S(k,l)$, which is unknown prior and it is the estimation target of the beamformer. An estimate of $S(k,l)$ is denoted by $\widehat{S}(k,l)$. A Lagrange multiplier method can be used to solve this optimization problem, and the cost function can be written as

$$\mathcal{J}_k = \sum_{l=1}^{\mathcal{L}} g(\mathbf{w}(k),\lambda_s(k,l)) + \alpha_k(\mathbf{w}^H(k)\mathbf{h}(k)-1), \tag{13}$$

where

$$\begin{aligned}&g(\mathbf{w}\left(k\right),\lambda_s\left(k,l\right))\\ &=\left(\frac{\left|\widehat{S}\left(k,l\right)\right|^2}{\lambda_s\left(k,l\right)}+\log(\pi\lambda_s\left(k,l\right))-\log\psi\left(\hat{\lambda}_s\left(k,l\right)\right)\right)\\ &=\left(\frac{\left|\mathbf{w}^{\mathrm{H}}\left(k\right)\mathbf{y}\left(k,l\right)\right|^2}{\lambda_s\left(k,l\right)}+\log\left(\pi\lambda_s\left(k,l\right)\right)-\log\psi\left(\lambda_s\left(k,l\right)\right)\right)\end{aligned} \tag{14}$$

and α_k denotes the Lagrange multiplier. This optimization problem needs to optimize $\lambda_s(k,l)$ and $\mathbf{w}(k)$ simultaneously, resulting in that we cannot get a closed-form solution. In this paper, we design an iterative optimization algorithm to solve this problem and the updating rule can be obtained by setting the partial differentials of the cost function with respect to a corresponding parameter at zero. When $\mathbf{w}(k)$ is determined, by setting the partial differentials of $\mathcal{J}_k$ with respect to $\lambda_s(k,l)$ at zero, one can get

$$\lambda_s(k,l) = \frac{2\gamma^{p/2}}{p}\left|\widehat{S}(k,l)\right|^{2-p}, \tag{15}$$

when $\lambda_s(k,l)$ is determined, the term related to $\lambda_s(k,l)$ can be regarded as a constant value, then the cost function (13) is equivalent to

$$\mathcal{J}_k=\sum_{l=1}^{\mathcal{L}} \frac{\left|\mathbf{w}^{\mathrm{H}}\left(k\right)\mathbf{y}\left(k,l\right)\right|^2}{\hat{\lambda}_s\left(k,l\right)}+\alpha_k\left(\mathbf{w}^{\mathrm{H}}\left(k\right)\mathbf{h}\left(k\right)-1\right). \tag{16}$$

Finally, by setting the partial differentials of (16) with repect to $\mathbf{w}(k)$ at zero and using the distortionless constraint on the desired speech, one can get

$$\widehat{\mathbf{w}}_{\mathrm{CGGD}}(k) = \frac{\left(_{\mathrm{CGGD}}\widehat{\mathbf{R}}_{\mathbf{yy}}(k)\right)^{-1}\mathbf{h}(k)}{\mathbf{h}^H(k)\left(_{\mathrm{CGGD}}\widehat{\mathbf{R}}_{\mathbf{yy}}(k)\right)^{-1}\mathbf{h}(k)}, \tag{17}$$

where

$$_{\mathrm{CGGD}}\widehat{\mathbf{R}}_{\mathbf{yy}}(k) = \sum_{l=1}^{\mathcal{L}} \frac{\mathbf{y}(k,l)\mathbf{y}^H(k,l)}{\lambda_s(k,l)} = \sum_{l=1}^{\mathcal{L}} \frac{p\mathbf{y}(k,l)\mathbf{y}^H(k,l)}{2\gamma^{p/2}\left|\widehat{S}(k,l)\right|^{2-p}}. \tag{18}$$

Because $\widehat{\mathbf{w}}_{\mathrm{CGGD}}(k)$ is invariant to the constant scaling factor in (15), (18) can be further reduced to

$$_{\mathrm{CGGD}}\widehat{\mathbf{R}}_{\mathbf{yy}}(k) = \sum_{l=1}^{\mathcal{L}} \frac{\mathbf{y}(k,l)\mathbf{y}^H(k,l)}{\max\left(\widehat{\lambda}_s^{(1-p/2)}(k,l),\delta\right)}, \tag{19}$$

where $\widehat{\lambda}_s(k,l) = |\widehat{S}(k,l)|^2$ denotes the estimated PSD of the desired speech in each time-frequency bin. δ in the denominator is a small positive floor value to avoid dividing by zero. In the iterative optimization algorithm, with the initialization $\widehat{\mathbf{w}}^0_{\mathrm{CGGD}}(k) = \widehat{\mathbf{w}}_{\mathrm{MPDR}}(k)$, we can keep updating $\widehat{\lambda}_s(k,l)$ and $\widehat{\mathbf{w}}_{\mathrm{CGGD}}(k)$ until reaching the maximum number of iterations.

B. Steering Vector Estimation

This part presents a steering vector estimation method in the iteration framework of the proposed CGGD-MLDR beamformer without explicitly estimating the desired speech PSD matrix. This paper adopts the generalized eigenvalue decomposition (GEVD)-based method for its stable performance [28], although both the noisy PSD matrix and the noise PSD matrix need to be estimated beforehand. To estimate the noise PSD matrix, we commonly need to estimate a time-frequency mask to distinguish noise-only bins from noisy bins, and then we can have

$$\widehat{\mathbf{R}}_{\mathbf{vv}}(k) = \frac{1}{\sum_{l=1}^{\mathcal{L}}\mathcal{M}(k,l)}\sum_{l=1}^{\mathcal{L}}\mathcal{M}(k,l)\mathbf{y}\left(k,l\right)\mathbf{y}^H\left(k,l\right), \tag{20}$$

where $\mathcal{M}(k,l) \in [0\ 1]$ denotes the time-frequency mask, which can be the speech absence probability. $\mathcal{M}(k,l)=1$ indicates that the time-frequency bin (k,l) is noise-only. On the contrary, $\mathcal{M}(k,l)=0$ means that the time-frequency bin (k,l) contains the desired speech. Traditionally, one needs to estimate the SPP, such as using complex Gaussian mixture model [29], to get this mask indirectly. It is interesting to see that (19) has the similar form as (20), and a larger $1/\widehat{\lambda}_s^{(1-p/2)}\left(k,l\right)$ implies that the desired speech PSD is lower in the time-frequency bin (k,l). This means that $1/\widehat{\lambda}_s^{(1-p/2)}\left(k,l\right)$ can be potentially applied to weight the sample covariance matrix, although it dose not range from zero to one. Accordingly, in this paper, we propose an alternative mask, which is

$$\mathcal{M}(k,l) = 1/\widehat{\lambda}_s^{(1-p/2)}\left(k,l\right). \tag{21}$$

Substitute (21) into (20) to obtain the estimated noise PSD matrix, and then perform the GVED-based method to estimate

Algorithm 1 CGGD-MLDR with SV estimation method

Input: $\mathbf{y}(k,l)$, p, $\hat{\mathbf{h}}$ and maximum iteration number I
Output: $\widehat{\mathbf{w}}^{I}_{\text{CGGD}}(k)$ and $\widehat{S}^{I}(k,l)$
Initialize: $\widehat{\mathbf{w}}^{0}_{\text{CGGD}}(k) = \widehat{\mathbf{w}}_{\text{MPDR}}(k)$
for $i = 0, 1, 2, ..., I-1$ **do**
  for $l = 1, 2, ..., \mathcal{L}$ **do**
    Compute $\widehat{S}^{i}(k,l) = (\widehat{\mathbf{w}}^{i}_{\text{CGGD}}(k))^{H}\mathbf{y}(k,l)$
    Update $\widehat{\lambda}^{i+1}_{s}(k,l) = \left|\widehat{S}^{i}(k,l)\right|^{2-p}$
    Compute
    ${}_{\text{CGGD}}\widehat{\mathbf{R}}^{i+1}_{\mathbf{yy}}(k) = {}_{\text{CGGD}}\widehat{\mathbf{R}}^{i+1}_{\mathbf{yy}}(k) + \frac{\mathbf{y}(k,l)\mathbf{y}^{H}(k,l)}{\widehat{\lambda}^{i+1}_{s}(k,l)}$
  Update $\widehat{\mathbf{w}}^{i+1}_{\text{CGGD}}(k) = \frac{\left({}_{\text{CGGD}}\widehat{\mathbf{R}}^{i+1}_{\mathbf{yy}}(k)\right)^{-1}\hat{\mathbf{h}}(k)}{\hat{\mathbf{h}}^{H}(k)\left({}_{\text{CGGD}}\widehat{\mathbf{R}}^{i+1}_{\mathbf{yy}}(k)\right)^{-1}\hat{\mathbf{h}}(k)}$
  Compute
  $\widehat{\mathbf{R}}_{\mathbf{vv}}(k) = \frac{1}{\sum_{l=1}^{\mathcal{L}}\mathcal{M}(k,l)}\sum_{l=1}^{\mathcal{L}}\mathcal{M}(k,l)\mathbf{y}(k,l)\,\mathbf{y}^{H}(k,l)$
  Update $\hat{\mathbf{h}}(k) = \widehat{\mathbf{R}}_{\mathbf{vv}}(k)\mathbb{P}\left\{\widehat{\mathbf{R}}^{-1}_{\mathbf{vv}}(k)\mathbf{R}_{\mathbf{yy}}(k)\right\}$
return $\widehat{\mathbf{w}}^{I}_{\text{CGGD}}(k)$ and $\widehat{S}^{I}(k,l)$

the steering vector, which can be given by:

$$\widehat{\mathbf{h}}(k) = \widehat{\mathbf{R}}_{\mathbf{vv}}(k)\mathbb{P}\left\{\widehat{\mathbf{R}}^{-1}_{\mathbf{vv}}(k)\mathbf{R}_{\mathbf{yy}}(k)\right\}, \tag{22}$$

where $\mathbb{P}\{\bullet\}$ extracts the principal eigenvector of a matrix.

Estimation of the steering vector and that of the weight vector for the CGGD-MLDR beamformer can be updated alternately to provide a very robust adaptive beamformer with a given coarse steering vector estimation and microphone array signals directly. The whole algorithm is summarized in Algorithm 1.

C. Related to the MPDR Beamfomer and its Variations

This part discuss the relationship between the proposed CGGD-MLDR beamformer and the MPDR beamformer together with its improved variations. The proposed CGGD-MLDR beamformer is a generalized MPDR beamformer, which can reduce to many existing improved versions of the MPDR beamformer for different values of the shape parameter p in (9). Obviously, we can have the following comments:

1) When $p = 2$, the proposed CGGD-MLDR beamformer reduces to the well-known MPDR beamformer due to ${}_{\text{CGGD}}\widehat{\mathbf{R}}_{\mathbf{yy}}(k) \equiv \widehat{\mathbf{R}}_{\mathbf{yy}}(k)$ because of $\widehat{\lambda}^{(1-p/2)}_{s}(k,l) \equiv 1$.
2) When $p = 0$, the proposed CGGD-MLDR beamformer becomes the newly proposed MLDR beamformer[1] [23] due to ${}_{\text{CGGD}}\widehat{\mathbf{R}}_{\mathbf{yy}}(k) \equiv{}_{\text{CGD}}\widehat{\mathbf{R}}_{\mathbf{yy}}(k)$ for $p = 0$.
3) When p is a positive value, for narrowband applications, the proposed CGGD-MLDR beamformer has the same form as the MDDR beamformer that is derived from the ℓ_p-norm [24]. Note that the proposed CGGD-MLDR beamformer gives clear guidelines on choosing the shape parameter p in (9), this is because it is derived from maximum likelihood theory and p relates to the distribution of the desired speech $S(k,l)$.

[1]In [30], the MLDR was also called the weighted MPDR (wMPDR).

D. Mechanisms of the Proposed CGGD-MLDR Beamformer in Improving Robustness

We assume that there are $\mathcal{L}_1$ noise-only frames among all $\mathcal{L}$ frames. We further assume that $\mathcal{L}$ and $\mathcal{L}_1$ is large enough that the intercorrelation between $\mathbf{x}(k,l)$ and $\mathbf{v}(k,l)$ can be ignored and $\lambda_s(k,l)$ and $\lambda_v(k,l)$ do not change over time. Accordingly, (19) becomes

$$\begin{aligned}{}_{\text{CGGD}}\mathbf{R}_{\mathbf{yy}}(k) &= \mathcal{L}_2\left(\lambda_s(k)\right)^{\frac{p}{2}}\boldsymbol{\Upsilon}_{ss}(k) \\ &\quad + \left(\mathcal{L}_1\rho(k) + \mathcal{L}_2\left(\lambda_s(k)\right)^{\frac{p}{2}}\frac{1}{\varepsilon(k)}\right)\boldsymbol{\Upsilon}_{vv}(k),\end{aligned} \tag{23}$$

where $\rho(k) = \lambda_v(k)/\delta^{(1-p/2)}$ and $\varepsilon(k) = \lambda_s(k)/\lambda_v(k)$ is the input SINR. $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ with $\mathcal{L}_2$ the number of frames containing desired speech. One can see that ${}_{\text{CGGD}}\mathbf{R}_{\mathbf{yy}}(k)$ is a linear combination of $\boldsymbol{\Upsilon}_{ss}(k)$ and $\boldsymbol{\Upsilon}_{vv}(k)$, we further define the ratio of the two combination coefficients, which is

$$r_p(k) = \frac{\mathcal{L}_1\rho(k) + \mathcal{L}_2\left(\lambda_s(k)\right)^{\frac{p}{2}}\frac{1}{\varepsilon(k)}}{\mathcal{L}_2\left(\lambda_s(k)\right)^{\frac{p}{2}}}. \tag{24}$$

when $p = 2$, we have $r_2(k) = \mathcal{L}/(\mathcal{L}_2\varepsilon(k))$ and thus $r_2(k)$ is determined by the input SINR and the number of the desired speech frames among all $\mathcal{L}$ noisy frames. Note that the smaller $r_2(k)$ is, the more sensitive the MPDR beamformer is. When $p = 0$, we have $r_0(k) = (\mathcal{L}_1\rho(k) + \mathcal{L}_2/\varepsilon(k))/\mathcal{L}_2$. Obviously, when $\rho(k)\varepsilon(k) \geq 1$, i.e., $\lambda_s(k) \geq \delta$, $r_0(k) \geq r_2(k)$ holds true always. This should be the reason that the MLDR beamformer can improve the robustness of the MDPR beamformer. For arbitrary values of $p \in [0\ 2)$, one can easily derive that $r_p(k) \geq r_2(k)$ holds true if and only if $\lambda_s(k) \geq \delta$, which means that the CGGD-MLDR beamformer can be always more robust than the MPDR beamformer due to that δ is only a small positive value as mentioned above.

IV. Experimental Results

This section evaluates the performance of the proposed CGGD-MLDR beamformer and compares it with the MPDR beamformer and the MLDR beamformer. The performance of the oracle minimum variance distortionless response (MVDR) beamformer is also presented to show the theoretical limit, where the noise PSD matrix and the steering vector of the desired speech are assumed to be known exactly for MVDR. In this section, the PESQ [31] and ESTOI [32] improvements are chosen as objective measurements.

A. Simulation Results

Ten utterances with each about 20s duration are taken from TIMIT corpus [33] and the white Gaussian noise is chosen from NOISEX-92 database [34] as interference. The room impulse response is generated by using the image method [35], with a room of size $6m \times 10m \times 4m$ and the reverberation time 160 ms. We consider a uniform linear array with 6 microphones and 4 cm inter-sensor distance which is placed at the center of the room. The desired speech is $2m$ away from the array center propagating from $\theta = 0°$, and two interferences propagate from $30°$ and $-60°$, respectively.

TABLE I
PESQ AND ESTOI IMPROVEMENTS OF CHiME-3 DATABASE. BEST SCORES ARE HIGHLIGHTED IN **BOLD**

Method	BUS		CAF		PED		STR	
	ΔPESQ	ΔESTOI	ΔPESQ	ΔESTOI	ΔPESQ	ΔESTOI	ΔPESQ	ΔESTOI
MPDR	-0.25	-0.05	0.31	0.09	0.34	0.11	0.05	0.01
MLDR	0.55	0.16	0.63	0.21	0.61	0.21	0.80	0.20
CGGD-MLDR	**0.65**	**0.17**	**0.68**	**0.23**	**0.65**	**0.26**	**0.88**	**0.22**

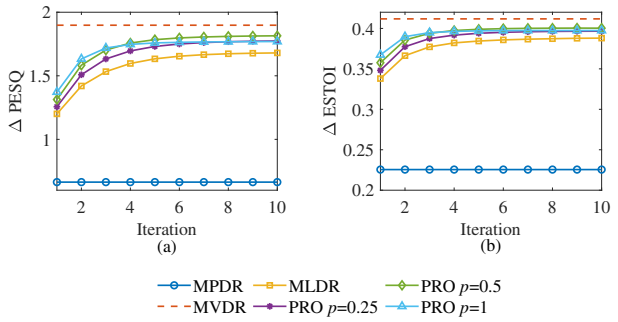

Fig. 1. Performance versus the number of iterations. (a) PESQ improvements. (b) ESTOI improvements. "PRO" represents "CGGD-MLDR" for simplicity.

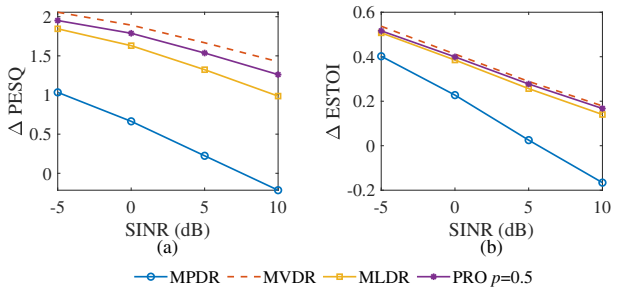

Fig. 2. Performance versus SINR. (a) PESQ improvements. (b) ESTOI improvements. "PRO" represents "CGGD-MLDR" for simplicity.

The steering vector of the desired speech is initialized by $\hat{\mathbf{h}}(k) = [1, 1, ..., 1]^T$.

In the first experiment, we study the performance of CGGD-MLDR beamformer versus the number of iterations with the input SINR=0 dB, where the results are plotted in Fig. 1. From this figure, one can observe that the choice of p seriously affect the performance of the CGGD-MLDR beamformer. Among them, it has the highest PESQ and ESTOI improvements with $p = 0.5$ and gradually converges to the theoretical limit with the increasing of the number of iterations, while MPDR have the lowest PESQ and ESTOI improvements. Moreover, one can see that the CGGD-MLDR beamformer with $p = 0.5$ provides much higher PESQ improvements with only very limited number of iterations, e.g., 2 to 3, which means much faster convergence rate than the MLDR beamformer.

In the following experiments, we only present the results of the CGGD-MLDR beamformer with $p = 0.5$ and compare it with the MLDR and the MPDR beamformers for its best performance in the first experiment. Fig. 2 plots the PESQ and ESTOI improvements versus the input SINR ranging from -5 dB to 10 dB. We can see that the PESQ improvements reduce as the increase of the input SINR for all beamformers and the proposed CGGD-MLDR beamformer with $p = 0.5$ is much better than the other two beamformers. For the MPDR beamformer, PESQ and ESTOI improvements can be negative because of the desired speech cancellation problem in high input SINR conditions.

B. CHiME-3 database Experiment results

Finally, we test the performance of the proposed CGGD-MLDR beamformer under different noisy scenarios using the CHiME-3 database. Experimental results in Table I are the average scores with 80 utterances in each scenario. One can also see that the proposed CGGD-MLDR beamformer with $p = 0.5$ can achieve the highest PESQ and ESTOI improvements under all noise scenarios. This demonstrates that the proposed CGGD-MLDR beamformer is robust to different types of noise and can still perform very well in real-world environments.

V. CONCLUSIONS

When a zero-mean complex generalized Gaussian distribution is introduced to model the complex spectrum of the desired speech, we derive the CGGD-MLDR beamformer with the maximum likelihood criterion, which is a generalization of the MPDR and the MLDR beamformers. Meanwhile, we propose to estimate the steering vector of the desired speech using the GEVD-based method with the help of the estimated speech PSD, and thus we do not need to estimate the speech presence probability and/or the speech absence probability. By properly choosing the shape parameter p, the CGGD-MLDR beamformer can achieve better performance than the MLDR beamformer in terms of PESQ and ESTOI improvements. The most attractive aspect is that the proposed CGGD-MLDR with $p = 0.5$ can converge in a much faster way than the MLDR beamformer, and so the computational complexity can be decreased dramatically due to that the proposed beamformer needs much fewer iterations to achieve the same performance of the MLDR beamformer.

REFERENCES

[1] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[2] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in israel*, pp. 416–420, 2010.

[3] W. Meng, Y. Ke, J. Li, C. Zheng and X. Li. "Finite data performance analysis of one-bit MVDR and phase-only MVDR," *Signal Processing* vol. 183, 108018, 2021.

[4] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[5] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, 2004.

[6] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 17, no. 7, pp. 1420–1434, 2009.

[7] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[8] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2009.

[9] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using maxnsr blocking matrix," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1493–1508, 2015.

[10] C. Zheng, A. Deleforge, X. Li, and W. Kellermann, "Statistical analysis of the multichannel wiener filter using a bivariate normal distribution for sample covariance matrices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 951–966, 2018.

[11] J. Benesty, C. Paleologu, C.-C. Oprea, and S. Ciochina, "An iterative multichannel wiener filter based on a kronecker product decomposition," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 211–215, 2020.

[12] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[13] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[14] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 67–79, 2013.

[15] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2499–2503, 2018.

[16] Y. Hu, T. D. Abhayapala, P. N. Samarasinghe, and S. Gannot, "Decoupled direction-of-arrival estimations using relative harmonic coefficients," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 246–250, 2020.

[17] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 223–233, 2011.

[18] J. H. Ko, J. Fromm, M. Philipose, I. Tashev, and S. Zarar, "Limiting numerical precision of neural networks to achieve real-time voice activity detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2236–2240, 2018.

[19] Y. Gu and A. Leshem, "Robust adaptive beamforming based on interference covariance matrix reconstruction and steering vector estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3881–3885, 2012.

[20] M. Taseska and E. A. P. Habets, "Nonstationary noise PSD matrix estimation for multichannel blind speech extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2223–2236, 2017.

[21] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1136–1150, 2019.

[22] C. Pan, J. Chen, and G. Shi, "On estimation of time-varying variances of source and noise for sensor array processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2865–2879, 2020.

[23] B. J. Cho, J.-M. Lee, and H.-M. Park, "A beamforming algorithm based on maximum likelihood of a complex gaussian distribution with time-varying variances for robust speech recognition," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1398–1402, 2019.

[24] X. Jiang, W.-J. Zeng, A. Yasotharan, H. C. So, and T. Kirubarajan, "Minimum dispersion beamforming for non-Gaussian signals," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1879–1893, 2014.

[25] L. Zhang, B. Li, L. Huang, T. Kirubarajan, and H. C. So, "Robust minimum dispersion distortionless response beamforming against fast-moving interferences," *Signal Processing*, vol. 140, pp. 190–197, 2017.

[26] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[27] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, 2015.

[28] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 544-548, 2015.

[29] T. Higuchi, N. Ito, T. Yoshioka and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5210-5214, 2016.

[30] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2267–2282, 2020.

[31] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, pp. 749–752, 2001.

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

[33] V. Zue, S. Seneff, J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech communication*, vol. 9, no. 4, pp. 351-356, 1990.

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

[35] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *The Journal of the Acoustical Society of America* , vol. 65, no. 3, pp. 943-950, 1979.