# Speech Enhancement via Maximum Likelihood Modal Beamforming with Complex Gaussian and Laplacian Priors

Shekhar Kumar Yadav, Graduate Student Member, IEEE and Nithin V. George, Member, IEEE
*Department of Electrical Engineering, Indian Institute of Technology Gandhinagar, India*
{yadav_shekhar, nithin}@iitgn.ac.in

***Abstract*—Capturing speech from a target source in the presence of interfering sources and noise using acoustic beamforming is an important processing tool for machine listening. When the target speaker can be at any location in the 3D space, spherical microphone arrays are desirable due to their ability to steer the beamformer towards any direction without affecting its directivity pattern. In this work, two beamformers based on the maximum likelihood estimation principle are introduced in the spherical harmonics domain. The first beamformer is designed by assuming that the coefficients of the target speech in the time-frequency domain at the beamformer output follow a zero-mean complex Gaussian prior distribution with time-varying variances. As speech coefficients are better modelled using Laplacian distribution, the second beamformer is designed by assuming a Laplacian prior for the target speech coefficients. To aid the capture of the desired speech, a distortionless constraint is also added to the formulation of both beamformers. The iterative update rules for the variances and the weights of both beamformers have been derived. Simulation results show that the proposed beamformers are more effective in target speech enhancement and distant speech recognition applications.**

***Index Terms*—Spherical microphone arrays, Acoustic beamforming, Speech enhancement, Maximum likelihood estimation**

## I. Introduction

In the presence of concurrent speech from interfering speakers and background noise, the ability to capture the desired speech while cancelling the interfering speeches and noise leads to target speech enhancement. It has applications in automatic distant speech recognition, robot audition, speech separation and diarization, human-robot interaction, binaural hearing aids etc. If the target speaker and the interfering speakers are at different spatial locations, then one of the best ways to reject the interfering speeches is by using acoustic beamforming [1]. The widely used MPDR (minimum power distortionless response) beamformer [2] was not specifically designed for speech signals. Recently, based on the assumption that the target speech in the time-frequency (TF) domain can be modelled using sparse priors such as a complex Gaussian distribution with time-varying variances, various speech processing algorithms have been developed [3]–[9]. However, most of these techniques, including acoustic beamforming and speech recognition, are developed for processing in the spatial domain. Spherical microphone arrays (SMAs), on the other hand, can take advantage of phase-mode processing by transforming the microphone signals into the spherical harmonics (SH) domain [10]–[12]. Array processing in the modal domain is advantageous because it leads to dimensionality reduction and a straightforward separation of the frequency parameters, array parameters and signal parameters. SMAs can also cover the entire 3D space without any directional ambiguity and their resolution is direction independent which is desirable in acoustic beamforming as the target source can be incident from any direction. Consequently, in recent years, there has been an increased focus on SH domain array processing [13]–[17].

In this work, we first introduce a beamformer in the SH domain that operates under the assumption that the target signal at the beamformer output follows a complex Gaussian distribution with time-varying variances. The weights of the beamformer are obtained based on the maximum likelihood estimation (MLE) principle. In order to further preserve the target speech, a distortionless constraint is added to the beamformer. Secondly, another beamformer in the SH domain is introduced which assumes a Laplacian prior on the target speech as it has been shown that speech coefficients in the TF domain are more accurately modelled using the Laplacian distribution rather than the complex Gaussian distribution [18]–[22]. The update rules for estimating the variances and the weights of both beamformers are derived in this work. Simulation results reveal the superiority of the proposed beamformers in speech enhancement and distant speech recognition applications.

*Notations:* $(.)^*, (.)^T, (.)^H$ are the conjugate, the non-conjugate transpose and the conjugate transpose operator, respectively. $\|.\|_p$ denotes the $\ell_p$-norm of a vector. $\mathbb{R}$ is the set of real numbers and $\mathbf{I}_M$ refers to an identity matrix of size $M \times M$.

## II. Preliminaries

We briefly introduce the SH domain signal model for an SMA placed in a sound field in this section. For a detailed discussion of SH signal model, readers are referred to [11]. We consider an SMA with $Q$ microphones on its surface of radius $r$. Let $\mathbf{r}_q$, $q \in \{1, 2, \ldots, Q\}$, be the position vector of the $q^{th}$ microphone (assuming the center of the SMA is at the origin). We assume that there are $D$ speakers located in the far-field of the SMA at locations $\Omega_d = (\theta_d, \phi_d)$, $d \in \{1, 2, \ldots, D\}$. Here,

This work is supported by the Department of Science and Technology, Government of India under the Core Grant Scheme (CRG/2018/002919) and TEOCO Chair of Indian Institute of Technology Gandhinagar.

$\theta_d$ refers to the elevation angle and $\phi_d$ refers to the azimuth angle of the $d^{th}$ source with respect to the centre of the SMA. In the Short Time Fourier Transform (STFT) domain, we can express the sound pressure at the $Q$ microphones at the $\tau^{th}$ time frame and $\nu^{th}$ frequency bin as [11]

$$\mathbf{p}(\tau,\nu) = \sum\nolimits_{d=1}^{D} \mathbf{h}(\nu,\Omega_d) S_d(\tau,\nu) + \boldsymbol{\eta}(\tau,\nu), \quad (1)$$

where $\mathbf{h}(\nu,\Omega_d) = [H(\nu,\Omega_d,\mathbf{r}_1),\ldots,H(\nu,\Omega_d,\mathbf{r}_Q)]^T$ and $H(\nu,\Omega_d,\mathbf{r}_q)$ represents the sound pressure at point $\mathbf{r}_q$ from a unit amplitude plane-wave arriving from $\Omega_d$. $S_d$ is the amplitude of the $d^{th}$ source and vector $\boldsymbol{\eta}$ represents the sensor noise at the $Q$ microphones. We process the microphone signals in the STFT domain as speech signals are non-stationary in nature. The SH transform $\mathbf{h}_{lm}(\nu,\Omega_d)$ of $\mathbf{h}(\nu,\Omega_d)$ can be expressed as [11], [23]

$$\mathbf{h}_{lm}(\nu,\Omega_d) = [b_l(k) Y^*_{lm}(\Omega_d)]_{l\geq 0, |m|<l}\,, \quad (2)$$

where $Y_{lm}$ is the SH of order $l$ and degree $m$ defined as

$$Y_{lm}(\Omega) = Y_{lm}(\theta,\phi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} \mathcal{P}_{lm}(\cos\theta) e^{\mathrm{j}m\phi}, \quad l\geq 0, |m|<l, \quad (3)$$

where $\mathcal{P}_{lm}$ is the associated Legendre function. In (2), $b_l(k)$ is referred to as the frequency-dependent mode strength which depends on various properties of the SMA such as microphone type, sampling locations, radius etc. and $k = \frac{2\pi\nu f_s}{cF}$ is the wavenumber ($c$ is the speed of sound, $f_s$ is the sampling frequency and $F$ is the number of frequency bins). Usually, it is convenient to express the SH signal model in a general form that is independent of the array properties. To cancel the dependency on array properties, we divide (2) by $b_l(k)$ which gives $\tilde{\mathbf{h}}_{lm}(\Omega_d) = \frac{\mathbf{h}_{lm}(\nu,\Omega_d)}{b_l(k)}$. Notice that $\tilde{\mathbf{h}}_{lm}(\Omega_d)$ is now independent of frequency and depends only on the direction-of-arrival angle $\Omega_d$. Now, we can express the spatial domain signal model (1) in the SH domain as

$$\mathbf{p}_{lm}(\tau,\nu) = \sum\nolimits_{d=1}^{D} \tilde{\mathbf{h}}_{lm}(\Omega_d) S_d(\tau,\nu) + \boldsymbol{\eta}_{lm}(\tau,\nu), \quad (4)$$

where $\mathbf{p}_{lm}$ and $\boldsymbol{\eta}_{lm}$ are referred to as the mode strength compensated SH transform of the spatial domain signal $\mathbf{p}$ and noise $\boldsymbol{\eta}$, respectively. Assuming aliasing free sampling [24] till $l = L$, we can express $\tilde{\mathbf{h}}_{lm}(\Omega_d)$ of size $((L+1)^2\times 1)$ as

$$\tilde{\mathbf{h}}_{lm}(\Omega_d) = [Y^*_{00}(\Omega_d), Y^*_{1(-1)}(\Omega_d), Y^*_{10}(\Omega_d),\ldots,Y^*_{LL}(\Omega_d)]^T.$$

In a reverberant room, the direct path sound pressure represented by $H(\nu,\Omega_d,\mathbf{r}_q)$ is substituted by the acoustic transfer function (ATF) $G(\nu,\Omega_d,\mathbf{r}_q)$ from the $d^{th}$ source to the microphone at $\mathbf{r}_q$. Consequently, for a reverberant environment, we obtain the SH signal model by replacing $\tilde{\mathbf{h}}_{lm}(\Omega_d)$ in (4) with the SH domain ATF vector $\tilde{\mathbf{g}}_{lm}(\Omega_d)$.

To perform beamforming and extract the desired speech $z(\tau,\nu)$ from $\mathbf{p}_{lm}(\tau,\nu)$ in each TF bin, the SH domain microphone signal is passed through a linear filter $\mathbf{w}_{lm}$ to get $z(\tau,\nu) = \mathbf{w}^H_{lm}(\nu)\mathbf{p}_{lm}(\tau,\nu)$. One of the most common beamformers in the modal domain is SH-MPDR [11] which operates by minimizing the beamformer output power while adding a distortionless constraint on the desired speech. Accordingly, in each $(\tau,\nu)$ TF bin, SH-MPDR is formulated as

$$\min_{\mathbf{w}_{lm}} \ \mathbf{w}^H_{lm}(\nu)\mathbf{R}_{\mathbf{p}_{lm}}\mathbf{w}_{lm}(\nu) \quad \text{s.t.} \quad \mathbf{w}^H_{lm}(\nu)\tilde{\mathbf{g}}_{lm}(\Omega_d) = 1. \quad (5)$$

where $\mathbf{R}_{\mathbf{p}_{lm}} = \mathbb{E}[\mathbf{p}_{lm}(\tau,\nu)\mathbf{p}^H_{lm}(\tau,\nu)]$ is the spatial covariance matrix of the microphone signal. The solution to (5) is given by $\mathbf{w}_{lm}(\nu) = \frac{(\mathbf{R}_{\mathbf{p}_{lm}})^{-1}\tilde{\mathbf{g}}_{lm}(\Omega_d)}{\tilde{\mathbf{g}}^H_{lm}(\Omega_d)(\mathbf{R}_{\mathbf{p}_{lm}})^{-1}\tilde{\mathbf{g}}_{lm}(\Omega_d)}$. After obtaining $z(\tau,\nu)$ using $\mathbf{w}_{lm}(\nu)$, we apply inverse STFT to get the estimated speech in the time domain. The objective is for the estimated speech to be as close to the desired speech as possible.

## III. Maximum Likelihood Distortionless Response Beamforming in the SH domain (SH-MLDR)

SH-MPDR does not assume that the target speech coefficients belong to any prior distribution in the STFT domain. In this section, we introduce two beamformers specifically designed for speech that maximize the likelihood that the desired STFT speech coefficients follow a suitable prior distribution.

*1) SH-MLDR with Complex Gaussian Prior:* In this subsection, we model the unknown STFT coefficients of the desired speech as a random variable having a circular zero-mean complex Gaussian (CG) prior with time and frequency dependent variances $\lambda(\tau,\nu)$ [4]. So, the probability density function (PDF) of the beamformer output can be expressed as

$$P(z(\tau,\nu)) = \frac{1}{\pi\lambda(\tau,\nu)} e^{-\frac{|z(\tau,\nu)|^2}{\lambda(\tau,\nu)}}. \quad (6)$$

It is assumed that $z(\tau_1,\nu_1)$ and $z(\tau_2,\nu_2)$ are independent for $(\tau_1,\nu_1)\neq(\tau_2,\nu_2)$. The unknown parameters to be estimated are the variances $\lambda(\tau,\nu)$ for all $(\tau,\nu)$ and the beamforming weights $\mathbf{w}_{lm}(\nu)$ for all $\nu$. Since the signal model in (4) and the PDF in (6) assume no dependency across frequencies, we can estimate the unknown parameters independently in each frequency bin by maximizing the likelihood function below

$$\mathcal{L}\left(\boldsymbol{\Theta}(\nu)\right) = \prod\nolimits_{\tau=1}^{T_\nu} P\left(z(\tau,\nu)\right), \quad (7)$$

where $\boldsymbol{\Theta}(\nu) = \{\mathbf{w}_{lm}(\nu), \lambda(1,\nu),\ldots,\lambda(T_\nu,\nu)\}$ is the set of unknown parameters and $T_\nu$ is the number of time frames in the $\nu^{th}$ frequency bin. To replace the product in (7) with a summation, we take its logarithm. Then, maximizing (7) is equivalent to minimizing the negative of the resulting log-likelihood function which is given as

$$\begin{aligned}\ell(\boldsymbol{\Theta}(\nu)) &= \sum_{\tau=1}^{T_\nu}\left(\log(|\lambda(\tau,\nu)|) + \frac{z(\tau,\nu)z^*(\tau,\nu)}{\lambda(\tau,\nu)}\right)\\ &= \sum_{\tau=1}^{T_\nu}\left(\log(|\lambda(\tau,\nu)|) + \frac{\mathbf{w}^H_{lm}\mathbf{p}_{lm}\mathbf{p}^H_{lm}\mathbf{w}_{lm}}{\lambda(\tau,\nu)}\right). \quad (8)\end{aligned}$$

Adding the distortionless constraint towards the target source direction $\mathbf{w}^H_{lm}\tilde{\mathbf{g}}(\Omega_d) = 1$ to (8) using the method of Lagrange multiplier results in an augmented cost function as

$$\begin{aligned}\mathcal{J}_\nu(\boldsymbol{\Theta}(\nu)) = \sum\nolimits_{\tau=1}^{T_\nu}&\left(\log(|\lambda(\tau,\nu)|) + \frac{\mathbf{w}^H_{lm}\mathbf{p}_{lm}\mathbf{p}^H_{lm}\mathbf{w}_{lm}}{\lambda(\tau,\nu)}\right)\\ &+ \beta_\nu(\mathbf{w}^H_{lm}\tilde{\mathbf{g}}_{lm}(\Omega_d) - 1), \quad (9)\end{aligned}$$

**Algorithm 1** SH-MLDR with Complex Gaussian Prior

**inputs**: $\mathbf{p}_{lm}(\tau,\nu)\ \forall\ \tau$, RTF estimate $\hat{\tilde{\mathbf{g}}}_{lm}(\Omega_d)$, $T_\nu$, $\epsilon_\nu$.
**initialize:** $\lambda(\tau,\nu) = |(\mathbf{w}^0_{lm}(\nu))^H \mathbf{p}_{lm}(\tau,\nu)|^2$
**repeat**
$\mathbf{w}_{lm}(\nu) \leftarrow$ use (12) with $\hat{\tilde{\mathbf{g}}}_{lm}(\Omega_d)$
$z(\tau,\nu) \leftarrow \mathbf{w}^H_{lm}(\nu)\mathbf{p}_{lm}(\tau,\nu)$
$\lambda(\tau,\nu) \leftarrow \max\{z(\tau,\nu)z^*(\tau,\nu), \epsilon_\nu\}$
**until**: condition satisfied
**output**: beamformer weights $\mathbf{w}_{lm}(\nu)$ for each $\nu$

**Algorithm 2** SH-MLDR with Laplacian Prior

**inputs**: $\mathbf{p}_{lm}(\tau,\nu)\ \forall\ \tau$, RTF estimate $\hat{\tilde{\mathbf{g}}}_{lm}(\Omega_d)$, $T_\nu$, $\epsilon_\nu$.
**initialize:** $\lambda(\tau,\nu) = |(\mathbf{w}^0_{lm}(\nu))^H \mathbf{p}_{lm}(\tau,\nu)|^2$
**repeat**
$\mathbf{w}_{lm}(\nu) \leftarrow$ solve (18) with $\hat{\tilde{\mathbf{g}}}_{lm}(\Omega_d)$
$z(\tau,\nu) \leftarrow \mathbf{w}^H_{lm}(\nu)\mathbf{p}_{lm}(\tau,\nu)$
$\lambda(\tau,\nu) \leftarrow \max\{(|\Re(z(\tau,\nu))| + |\Im(z(\tau,\nu))|)^2, \epsilon_\nu\}$
**until**: condition satisfied
**output**: beamformer weights $\mathbf{w}_{lm}(\nu)$ for each $\nu$

where $\beta_\nu$ is the Lagrange multiplier. Since the cost function depends on variables $\lambda(\tau,\nu)$ and $\mathbf{w}_{lm}(\nu)$, their optimal values cannot be obtained by minimizing $\mathcal{J}_\nu$ analytically. Rather, the variables have to be estimated using iterative update rules. We can find the update rule for each variable by taking the partial derivative of the cost function in (9) with respect to the corresponding variable and equating it to zero while assuming the other variable to be constant. So, the update rule for $\lambda(\tau,\nu)$ after setting the partial derivative to zero is

$$\frac{\partial \mathcal{J}_\nu(\mathbf{\Theta}(\nu))}{\partial \lambda(\tau,\nu)} = \frac{1}{\lambda(\tau,\nu)} - \frac{z(\tau,\nu)z^*(\tau,\nu)}{\lambda^2(\tau,\nu)} = 0,$$
$$\lambda(\tau,\nu) = z(\tau,\nu)z^*(\tau,\nu) = |\mathbf{w}^H_{lm}(\nu)\mathbf{p}_{lm}(\tau,\nu)|^2. \quad (10)$$

In a similar way, setting the partial derivative of $\mathcal{J}_\nu$ w.r.t $\mathbf{w}_{lm}(\nu)$ to zero gives

$$\sum_{\tau=1}^{T_\nu} \left( \frac{\mathbf{p}_{lm}\mathbf{p}^H_{lm}}{\lambda(\tau,\nu)} \right) \mathbf{w}_{lm}(\nu) + \beta_\nu \tilde{\mathbf{g}}_{lm}(\Omega_d) = \mathbf{0}. \quad (11)$$

Solving (11) after eliminating $\beta_\nu$ by assuming that the distortionless constraint is satisfied, the closed-form update rule for $\mathbf{w}_{lm}(\nu)$ is obtained as

$$\mathbf{w}_{lm}(\nu) = \frac{(\mathbf{R}_{\tilde{\mathbf{p}}_{lm}})^{-1} \tilde{\mathbf{g}}_{lm}(\Omega_d)}{\tilde{\mathbf{g}}^H_{lm}(\Omega_d) (\mathbf{R}_{\tilde{\mathbf{p}}_{lm}})^{-1} \tilde{\mathbf{g}}_{lm}(\Omega_d)}. \quad (12)$$

where $\mathbf{R}_{\tilde{\mathbf{p}}_{lm}} = \sum_{\tau=1}^{T_\nu} \left( \frac{\mathbf{p}_{lm}\mathbf{p}^H_{lm}}{\lambda(\tau,\nu)} \right)$. The two-step update process is repeated iteratively until convergence is reached or a fixed number of iterations is completed. The complete process of SH-MLDR beamformer with complex Gaussian prior (SH-MLDR-CG) is outlined in Algorithm 1. In the first iteration, weight $\mathbf{w}^0_{lm}(\nu)$ is initialized with the weight of SH-MPDR. $\epsilon_\nu$ is a small real-valued positive constant to avoid very small values of $\lambda(\tau,\nu)$ for robust estimation of $\mathbf{R}_{\tilde{\mathbf{p}}_{lm}}$.

*2) SH-MLDR with Laplacian Prior:* It has been reported in various works of literature that speech coefficients are better modelled using Laplacian distribution (LD) rather than using Gaussian distribution [19]. This is because LD leads to sparser coefficients as it is steeper at the mean and has a heavier tail. In this subsection, we introduce the SH-MLDR beamformer where the desired speech at the beamformer output is modelled locally in each TF bin with an LD. Both the real and imaginary components of $z(\tau,\nu)$ are assumed to be independently modelled with a zero-mean LD with equal variances of $\lambda(\tau,\nu)/2$. The PDF of the beamformer output is then given as

$$P\left(z(\tau,\nu)\right) = \frac{1}{\lambda(\tau,\nu)} e^{-2\frac{|\Re(z(\tau,\nu))| + |\Im(z(\tau,\nu))|}{\sqrt{\lambda(\tau,\nu)}}}, \quad (13)$$

where $\Re(.)$ and $\Im(.)$ are the real and imaginary components of a complex number, respectively. Now, similar to the discussion in the previous subsection, we can estimate the unknown parameters by minimizing the negative of the log-likelihood function of the joint PDF which is given as

$$\tilde{\ell}(\mathbf{\Theta}(\nu)) = \sum_{\tau=1}^{T_\nu} \left( \log\lambda(\tau,\nu) + 2\frac{|\Re(z(\tau,\nu))| + |\Im(z(\tau,\nu))|}{\sqrt{\lambda(\tau,\nu)}} \right), \quad (14)$$

(14) is minimized along with the distortionless constraint $\mathbf{w}^H_{lm}(\nu)\tilde{\mathbf{g}}_{lm}(\Omega_d) = 1$. The proposed cost function in (14) can be differentiated with respect to $\lambda(\tau,\nu)$. So, setting its partial derivative to zero, we get the update rule for $\lambda(\tau,\nu)$ as

$$\frac{\partial \tilde{\ell}(\mathbf{\Theta}(\nu))}{\partial \lambda(\tau,\nu)} = \frac{1}{\lambda(\tau,\nu)} - \frac{|\Re(z(\tau,\nu))| + |\Im(z(\tau,\nu))|}{\lambda^{\frac{3}{2}}(\tau,\nu)} = 0,$$
$$\lambda(\tau,\nu) = \left(|\Re(z(\tau,\nu))| + |\Im(z(\tau,\nu))|\right)^2. \quad (15)$$

The cost function in (14) along with the distortionless constraint is not differentiable with respect to $\mathbf{w}_{lm}(\nu)$. Rewriting the cost function in (14) in terms of $\mathbf{w}_{lm}$ gives

$$\tilde{\ell}(\mathbf{w}_{lm}) = \sum_{\tau=1}^{T_\nu} \frac{2}{\sqrt{\lambda(\tau,\nu)}} \Big( |\Re(\mathbf{w}^H_{lm}(\nu)\mathbf{p}_{lm}(\tau,\nu))| + |\Im(\mathbf{w}^H_{lm}(\nu)\mathbf{p}_{lm}(\tau,\nu))| \Big) + \gamma_\nu, \quad (16)$$

where $\gamma_\nu$ does not depend on $\mathbf{w}_{lm}(\nu)$. Now, the estimate of $\mathbf{w}_{lm}$ can be obtained as a solution to the problem below

$$\min_{\mathbf{w}_{lm}} \ \tilde{\ell}(\mathbf{w}_{lm}) \quad \text{s.t.} \quad \mathbf{w}^H_{lm}\tilde{\mathbf{g}}_{lm}(\Omega_d) = 1. \quad (17)$$

(17) can be solved using the CVX toolbox [25] by converting it into the following linear programming problem [26]

$$\begin{aligned} \min_{\mathbf{v},\mathbf{w}_{lm}} \quad & \|\mathbf{v}\|_1 \\ \text{s.t.} \quad & \mathbf{v} \geq 0 \\ & |\Re(\mathbf{w}^H_{lm}(\nu)\mathbf{p}_{lm}(\tau,\nu))| \leq \frac{\sqrt{\lambda(\tau,\nu)}}{2} v_{2\tau-1} \\ & |\Im(\mathbf{w}^H_{lm}(\nu)\mathbf{p}_{lm}(\tau,\nu))| \leq \frac{\sqrt{\lambda(\tau,\nu)}}{2} v_{2\tau} \\ & \mathbf{w}^H_{lm}\tilde{\mathbf{g}}_{lm}(\Omega_d) = 1, \end{aligned} \quad (18)$$

TABLE I

OBJECTIVE MEASURES FOR THE PERCEPTUAL EVALUATION OF THE TARGET SPEECH AT THE OUTPUT OF THE VARIOUS BEAMFORMER COMPARED TO THE CLEAN SPEECH IN THE TIME DOMAIN. THE NUMBER IN **D** REFERS TO THE DESIRED SPEAKER.

| | | Anechoic (AE) | | | | $RT_{60} = 150$ ms | | | | $RT_{60} = 200$ ms | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Beamformers** | **D** | **CD** | **PESQ** | **WSS** | **LLR** | **CD** | **PESQ** | **WSS** | **LLR** | **CD** | **PESQ** | **WSS** | **LLR** |
| **SH-MPDR** | 1 | 3.2851 | 1.3787 | 31.0631 | 0.3646 | 3.9914 | 1.2354 | 50.1410 | 0.5761 | 4.4721 | 1.1140 | 59.5806 | 0.6859 |
| | 2 | 3.1382 | 1.4382 | 27.1895 | 0.3256 | 3.6186 | 1.5653 | 47.1980 | 0.4988 | 3.9729 | 1.3922 | 58.4016 | 0.5561 |
| **SH-MLDR-CG** | 1 | 2.9094 | 1.9707 | 24.5401 | 0.3250 | 3.645 | 1.8754 | 45.4062 | 0.5137 | 4.2426 | 1.6885 | 53.1389 | 0.5496 |
| | 2 | 2.5565 | 2.1113 | 22.7916 | 0.2614 | 3.3643 | 1.9894 | 41.4946 | 0.4362 | 3.5262 | 1.5304 | 51.6684 | 0.4884 |
| **SH-MLDR-LP** | 1 | 2.7600 | 2.2622 | 22.3734 | 0.3022 | 3.6279 | 2.0331 | 42.7903 | 0.4493 | 3.8827 | 1.8195 | 49.6425 | 0.5050 |
| | 2 | 2.2839 | 2.2967 | 19.6472 | 0.2301 | 3.0189 | 2.1361 | 39.1988 | 0.385 | 3.4469 | 2.0151 | 47.497 | 0.4389 |

where variable $\mathbf{v} \in \mathbb{R}^{2T_\nu}$ and $v_\tau$ is the $\tau^{th}$ element of $\mathbf{v}$. The update of $\lambda(\tau, \nu)$ using (15) and update of $\mathbf{w}_{lm}$ using (18) is performed iteratively until some convergence criterion is reached or a certain number of iterations is completed. The procedure of SH-MLDR beamformer with Laplacian prior (SH-MLDR-LP) is summarized in Algorithm 2. Although SH-MLDR-LP requires solving a linear programming problem, it is expected to perform better than SH-MLDR-CG. This is because Laplacian prior leads to sparser solutions when compared to Gaussian prior which is desirable as the beamformer output should be sparser than the mixed speech input signal captured by the microphone array. The inducing of sparsity at the beamformer output can also be seen from the appearance of $\ell_1$-norm minimization in (18). Also, modern solvers dealing with linear programs have become fast enough and can be comfortably deployed in real-time applications.

## IV. PERFORMANCE EVALUATION

The acoustic beamforming performances of SH-MPDR, SH-MLDR-CG and SH-MLDR-LP are compared in this section. A 20 microphone SMA of radius 3 $cm$ at the centre of a room with dimensions $7 \times 5 \times 4$ $m^3$ is considered. The microphones are placed at the centre of the faces of a dodecahedron. We assume that two speech sources are located at $(45°, 30°)$ and $(135°, 30°)$ at a distance of 2 $m$ from the array. The anechoic speeches were taken from the standard Librispeech dataset [27] which is a corpus of audiobooks in the public domain. The sampling frequency was 16 kHz and the duration of each speech was 4 seconds. In the reverberant case, we generated the reverberant speeches by convolving both speeches with the room impulse response (RIR) from the locations of the sources to the microphones. The image method [28] was used to generate the RIRs. For the generation of the RIRs, we consider two cases: 1) reverberation time ($RT_{60}$) of 150 $ms$ and 2) $RT_{60}$ of 200 $ms$. While processing in the TF domain, the STFT window length was chosen to be 1024 which corresponds to 64 $ms$ of data in one time frame. Hanning window was used, and an overlap of 50% was set between consecutive frames ($N_\tau = 124$). Additive Gaussian noise with a signal-to-noise ratio of 20 dB was added to the captured microphone signals in all the simulations. We set the number of iterations for SH-MLDR-CG and SH-MLDR-LP to three. The RTF estimation necessary for all three

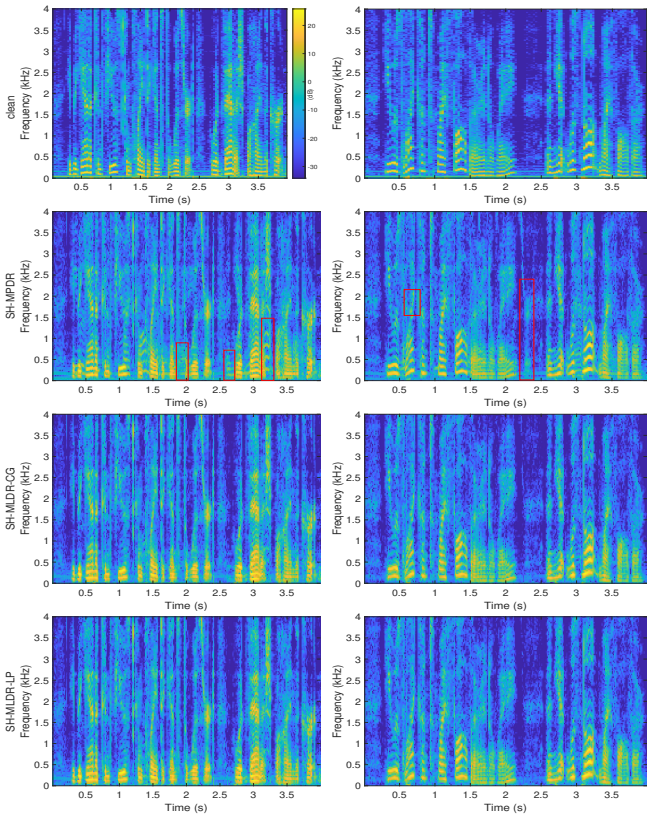


Fig. 1. Output STFT spectrograms (SG) of the beamformers. Interfering residuals are highlighted using the red boxes. Each column shows the SG when the distortionless constraint is towards one of the two speakers, respectively.

beamformers was done using the method laid out in [29]. Fig. 1 shows the spectrograms of the clean speeches as well as the spectrograms at the beamformer outputs for comparison for the anechoic case. We can see that SH-MPDR is unable to remove the interfering speech and noise residuals from the target speech, whereas both SH-MLDR-CG and SH-MLDR-LP are able to completely remove the residual elements from the target speech leading to significant speech enhancement. For more clarity on the performance, the output of the various beamformers was compared using objective measures such as Cepstral Distance (CD), Perceptual Evaluation of Speech Quality (PESQ), Weighted Slope Spectral (WSS) distance and Log-likelihood Ratio (LLR) [30]. The measures for each

TABLE II
COMPARISON OF DISTANT SPEECH RECOGNITION (DSR) PERFORMANCE

| | $RT_{60}$ | SH-MPDR | | SH-MLDR-CG | | SH-MLDR-LP | |
|---|---|---|---|---|---|---|---|
| **S** | | 1 | 2 | 1 | 2 | 1 | 2 |
| **W** | AE | 46.67 | 42.86 | 9.09 | 6.67 | 9.09 | 6.67 |
| **E** | 0.15 s | 57.14 | 46.67 | 33.33 | 26.67 | 21.43 | 21.43 |
| **R** | 0.2 s | 64.29 | 53.33 | 35.71 | 33.33 | 26.74 | 26.74 |
| **S** | AE | 6.078 | 8.731 | 12.184 | 13.164 | 13.27 | 14.521 |
| **D** | 0.15 s | 1.567 | 1.593 | 8.017 | 8.164 | 9.041 | 9.611 |
| **R** | 0.2 s | -1.420 | 0.263 | 5.103 | 5.841 | 6.288 | 6.615 |

beamformer are listed in Table I for both sources. It is clear from the table that SH-MLDR-LP performs better in separating and enhancing the desired speech than SH-MLDR-CG; however, both the proposed beamformers outperform SH-MPDR even in moderately reverberant environments. This is because, in the formulation of the proposed beamformers in (9) and (17), the presence of the inverse of the $\lambda(l, f)$ term plays a crucial role as it ensures that TF bins with low variance (possibly belonging to interfering speeches and noise component) are given a higher priority when minimizing the power at the output of the beamformer and TF bins with high variance (possibly containing content from the desired speech) are given a lower priority, which adds more credibility to the proposed beamformers than a conventional SH-MPDR beamformer which does not take into account the variance of the TF bins. One of the major applications of acoustic beamformers is distant speech recognition (DSR) in the presence of interfering speakers and isotropic noise. The comparison of DSR performance in terms of Signal-to-Distortion Ratio (SDR) and Word Error Rate (WER) is presented in Table II. Once again, it can be seen that SH-MLDR-LP performs the best and that both the proposed beamformers outperform the SH-MPDR beamformer.

## V. CONCLUSION

Maximum likelihood distortionless response (MLDR) beamformers with complex Gaussian prior (CGP) and Laplacian prior (LP) in the spherical harmonics (SH) domain are introduced in this work. The update rules for the variances and weights of the beamformer using CGP leads to closed-form solutions whereas using LP requires solving a linear program. However, the speech enhancement performance of SH-MLDR-LP is better than SH-MLDR-CG due to more accurate modelling of the target speech signal. In extensive future work, additional features will be added to the proposed beamformers so that they can handle high reverberation and more complicated acoustic environments, along with the development of an online version of the proposed beamformers.

## REFERENCES


[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[2] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[3] T. Nakatani *et al.*, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 2267–2282, 2020.

[4] 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 Process. Lett.*, vol. 26, no. 9, pp. 1398–1402, 2019.

[5] B. J. Cho and H.-M. Park, "Convolutional maximum-likelihood distortionless response beamforming with steering vector estimation for robust speech recognition," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 1352–1367, 2021.

[6] T. Nakatani *et al.*, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[7] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, 2019.

[8] ——, "Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation," in *27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[9] S. K. Yadav and N. V. George, "Distortionless acoustic beamforming with enhanced sparsity based on reweighted $\ell_1$-norm minimization," in *International Congress on Acoustics (ICA)*, 2022, pp. A05–227–234.

[10] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2018, vol. 16.

[11] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and applications of spherical microphone array processing*. Springer, 2017, vol. 9.

[12] J. R. Driscoll and D. M. Healy, "Computing fourier transforms and convolutions on the 2-sphere," *Advances in applied mathematics*, vol. 15, no. 2, pp. 202–250, 1994.

[13] V. Varanasi *et al.*, "A deep learning framework for robust DOA estimation using spherical harmonic decomposition," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 1248–1259, 2020.

[14] S. K. Yadav and N. V. George, "Coarray manifold separation in the spherical harmonics domain for enhanced source localization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 5038–5042.

[15] A. H. Moore and P. A. Naylor, "Linear prediction based dereverberation for spherical microphone arrays," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[16] J.-W. Choi *et al.*, "Multiarray eigenbeam-esprit for 3D sound source localization with multiple spherical microphone arrays," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 30, pp. 2310–2325, 2022.

[17] S. K. Yadav *et al.*, "Sparse distortionless modal beamforming for spherical microphone arrays," *IEEE Signal Process. Lett.*, pp. 1–5, 2022.

[18] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, 2003.

[19] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with laplacian model of the desired signal," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 5172–5176.

[20] A. Jukić *et al.*, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[21] M. Witkowski and K. Kowalczyk, "Split bregman approach to linear prediction based dereverberation with enforced speech sparsity," *IEEE Signal Process. Lett.*, vol. 28, pp. 942–946, 2021.

[22] M. Fraś and K. Kowalczyk, "Convolutional weighted parametric multichannel wiener filter for reverberant source separation," *IEEE Signal Process. Lett.*, vol. 29, pp. 1928–1932, 2022.

[23] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2149–2157, 2004.

[24] B. Rafaely *et al.*, "Spatial aliasing in spherical microphone arrays," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010, 2007.

[25] M. Grant and S. Boyd. CVX: MATLAB software. [Access on: Feb. 2023]. [Online]. Available: http://cvxr.com/cvx/

[26] S. Boyd *et al.*, *Convex optimization*. Cambridge university press, 2004.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[28] E. Habets. Room impulse response (RIR) generator. [Access on: Feb. 2023]. [Online]. Available: https://github.com/ehabets/RIR-Generator

[29] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[30] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.