25-29 August 2024
inter.noise
NANTES FRANCE

Acoustic Beamforming using Maximum Likelihood Criteria with Laplacian Model of the Desired Signal

Jovit Jayan[1], Kushal Rathod[2], Shekhar Kumar Yadav[3], Nithin V. George[4]
Department of Electrical Engineering, Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat 382355, India

ABSTRACT

Acoustic beamforming is a technique used in audio engineering and signal processing to filter sound waves in the spatial domain. Usually, an array of microphones is used to capture sound from different directions. These array signals are then combined to reinforce each other in the desired direction while cancelling the noise or interference from other directions. An important application of acoustic beamforming is the cocktail party scenario where multiple people speak simultaneously in a noisy room. To capture the speech of only the desired speaker, acoustic beamforming is used. Usually, in such cases, the target speech is modelled as a zero-mean complex Gaussian distributed random variable. However, the target speech coefficients are sparse in the time-frequency domain. Hence, in our work, we model the speech coefficients using a zero-mean circular Laplacian distribution. After modelling the target speech, we formulate a beamformer based on the maximum likelihood criteria. We add a distortionless constraint to the proposed beamformer to further improve performance. The final solution of the proposed beamformer encourages sparsity, indicating that it models the target speech better than the complex Gaussian distribution. Simulations show the effectiveness of the proposed beamformer in capturing the target speech and rejecting interfering speakers.

1. INTRODUCTION

In an acoustic scenario where multiple speakers at different locations are speaking at the same time, a single microphone will capture the speech of all the speakers simultaneously. However, using an array of microphones, only the speech of the desired speaker can be captured while filtering out the speech of the interfering speakers. This is known as acoustic beamforming, which is an acoustic spatial filtering technique [1]. Acoustic beamforming has many applications, such as machine listening, robot audition, hands-free telephony, smart home assistant devices, automatic speech recognition, hearing aids, virtual acoustics, etc. Acoustic beamforming is typically formulated by passing the microphone array signal through a spatial filter whose weights are adjusted based on certain criteria to obtain the desired speech at the beamformer output. One of the most common and practical beamformers is the minimum power distortionless response (MPDR) beamformer [2], which is formulated by minimizing the power of the beamformer output

[1]jovit.jayan@iitgn.ac.in (Equal contribution)
[2]kushal.rathod@iitgn.ac.in (Equal contribution)
[3]yadav_shekhar@iitgn.ac.in
[4]nithin@iitgn.ac.in

signal while pitting a distortionless constraint in the direction of the desired speaker. However, the drawback of the MPDR beamformer is that it is not specifically designed for speech applications. It does not take advantage of the different features of speech signals, which results in sub-par performances.

In recent years, numerous works in the field of acoustic denoising and dereverberation have modelled the desired speech coefficients in the short-time Fourier transform (STFT) domain as a zero-mean circular complex Gaussian random variable [3–7]. An acoustic beamformer was also recently designed by modelling the speech coefficients at the output of the beamformer in the STFT domain as a zero-mean circular complex Gaussian random variable [8, 9]. The variances of the prior distribution were designed to be time and frequency varying to reflect the non-stationary nature of speech signals in the STFT domain. The beamformer was named as the maximum likelihood distortionless response (MLDR) beamformer as the weight vector and the variances were estimated using a maximum likelihood technique.

The MLDR beamformer using the complex Gaussian prior does not take into account the sparse nature of a clean speech signal in the STFT domain. Modelling the desired speech coefficients with a prior distribution that promotes sparsity has been incorporated into various recent works [10–15]. In the case of acoustic beamforming, the input signal of the beamformer is dense in the STFT domain as it contains the speech coefficients of all the speakers. However, the output signal of the beamformer is sparse as it corresponds to the speech of only a single desired speaker. Hence, sparsifying the beamformer output should result in a better performance.

In this work, we design an acoustic beamformer by modelling the real and imaginary parts of the clean speech in the SFTF domain at the beamformer output as random variables that follow a zero-mean Laplacian distribution with time and frequency varying variances. The Laplacian distribution has a heavier tail than the complex Gaussian distribution (CG), and as such, it promotes more sparsity than the CG distribution. We also add the distortionless constraint in the formulation of the proposed beamformer. We utilize the maximum likelihood technique to obtain the beamforming weight vector and the estimates of the variances of the proposed beamformer. The weight vector of the proposed beamformer is obtained by solving a linear program, whereas a closed-form solution is derived for estimating the variances in the different time-frequency bins. Simulation results illustrate the effectiveness of the proposed beamformer in capturing the speech of the desired speaker while filtering out the speech of the interfering speakers as well as the microphone noise.

2. PRELIMINARIES

2.1. Microphone Array Signal Model

Let us consider that a microphone array with P microphones is placed in an anechoic or a low reverberant room. Let there be D speakers present in the room at different locations in the far field of the array. Let $\Omega_d = (\theta_d, \phi_d)$ represent the location of the d^{th} speaker, where θ_d and ϕ_d correspond to the elevation and azimuth angle of the speaker, respectively. In the STFT domain, the speech signal captured by the microphone array when all the D speakers are speaking simultaneously is given as

$$\mathbf{x}(t,f) = \sum_{d=1}^{D} \mathbf{a}(f,\Omega_d) S_d(f,t) + \mathbf{n}(t,f), \tag{1}$$

where $S_d(t,f)$ is the STFT speech coefficient of the d^{th} speaker in the t^{th} time frame and f^{th} frequency bin. $\mathbf{a}(f,\Omega_d)$ is the steering vector or the non-convolutive transfer function of the d^{th} speaker to the P microphones. In this work, we will deal with an anechoic or low reverberation environment, and as such, the transfer function is considered to be non-convolutive in nature. Further, in Equation 1, $\mathbf{n}(t,f)$ denotes the vector containing the noise component of the microphones.

The objective of the work in this paper is to capture the speech of only the desired speaker and filter out the speech of the interfering speakers. To do so, acoustic beamforming or spatial filtering is used. To perform acoustic beamforming, the array signal is passed through a spatial filter and the output of the filter is expressed as

$$z(t,f) = \mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f), \tag{2}$$

where $\mathbf{w}(f)$ denotes the filter weight in the f^{th} frequency bin. $(.)^{\mathrm{H}}$ denotes the complex conjugate operator. The beamformer output $z(t,f)$ should match the speech signal of the desired speaker $S_d(t,f)$. To obtain the desired speech in the time domain, we apply inverse STFT to signal at the beamformer output. We next discuss two common beamformers used for acoustic beamforming.

2.2. Minimum Power Distortionless Response (MPDR) Beamformer

The MPDR beamformer is formulated by minimizing the power of the beamformer output while putting a distortionless constraint in the direction of the desired speaker. Therefore, the weight vector of the MPDR beamformer is obtained by solving the following optimization problem

$$\min_{\mathbf{w}(f)} \quad \mathbf{w}^{\mathrm{H}}(f)\mathbf{R}_{\mathbf{x}}(f)\mathbf{w}(f) \qquad \text{s.t.} \qquad \mathbf{w}^{\mathrm{H}}(f)\mathbf{a}(f,\Omega_d) = 1, \tag{3}$$

where $\mathbf{R}_{\mathbf{x}}(f) = \frac{1}{N_t}\sum_{t=1}^{N_t}\mathbf{x}(t,f)\mathbf{x}^{\mathrm{H}}(t,f)$ is the array covariance matrix in the f^{th} frequency bin. N_t is the total number of time frames in a particular frequency bin. The distortionless constraint $\mathbf{w}^{\mathrm{H}}(f)\mathbf{a}(f,\Omega_d) = 1$ ensures that the speech signals coming from the desired direction are not distorted. Equation 3 has a closed-form solution, and consequently, the weight vector of the MPDR beamformer is given as

$$\mathbf{w}(f) = \frac{\mathbf{R}_{\mathbf{x}}^{-1}(f)\mathbf{a}(f,\Omega_d)}{\mathbf{a}^{\mathrm{H}}(f,\Omega_d)\mathbf{R}_{\mathbf{x}}^{-1}(f)\mathbf{a}(f,\Omega_d)}. \tag{4}$$

2.3. Maximum Likelihood Distortionless Response (MLDR) Beamformer with Complex Gaussian Prior

The MPDR beamformer is a very general-purpose beamformer. The drawback of using the MPDR beamformer for speech applications is that it does not take advantage of the specific features of speech signals and, as such, is not suitable for speech applications. In this subsection, we will discuss an acoustic beamformer that uses prior information about the distribution of speech signals. It is known that human speech coefficients in the STFT domain, when treated as a random variable, can be assumed to follow a zero-mean circular and complex Gaussian (CG) distribution. MLDR beamformer is designed such that the beamformer output follows a CG prior distribution with time and frequency varying variances. The variance of the prior distribution is time and frequency varying to account for the non-stationary nature of speech signals in the STFT domain. So, the probability distribution function (PDF) of the MLDR beamformer output is given as

$$\mathscr{P}(z(t,f)) = \frac{1}{\pi\gamma(t,f)}e^{-\frac{|z(t,f)|^2}{\gamma(t,f)}}, \tag{5}$$

where $\gamma(t,f)$ represents the variance of the prior distribution in the t^{th} time frame and f^{th} frequency bin. Since there is no dependency between the array signals across the different frequency bins, we can process signals in each frequency bin independently. Assuming that the signals in consecutive time frames are independent of each other, the joint PDF in a particular frequency bin can be expressed as

$$\mathscr{J}(\boldsymbol{\beta}_f) = \prod_{t=1}^{N_t}\mathscr{P}(z(t,f)), \tag{6}$$

where $\boldsymbol{\beta}_f = \{\mathbf{w}(f), \gamma(1,f), \gamma(2,f), \ldots, \gamma(N_t,f)\}$ is the set of unknown parameters that have to be estimated. The weight vector and the variances of the MLDR beamformer are obtained by maximizing the likelihood function of the PDF, which is the same as minimizing the negative of the log-likelihood function of the PDF, which is given as

$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta}_f) &= -\log(\mathscr{P}(z(t,f))) = \sum_{t=1}^{N_t}\left(\log(\pi) + \log(\gamma(t,f)) + \frac{|z(t,f)|^2}{\gamma(t,f)}\right) \\ &= \sum_{t=1}^{N_t}\left(\log(\pi) + \log(\gamma(t,f)) + \frac{\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f)\mathbf{x}^{\mathrm{H}}(t,f)\mathbf{w}(f)}{\gamma(t,f)}\right). \end{aligned} \tag{7}$$

Just like the MPDR beamformer, in the formulation of the MLDR beamformer, a distortionless constraint is added to the cost function in Equation 7 to get the final cost function as

$$\mathscr{C}(\boldsymbol{\beta}_f) = \sum_{t=1}^{N_t}\left(\log(\pi) + \log(\gamma(t,f)) + \frac{\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f)\mathbf{x}^{\mathrm{H}}(t,f)\mathbf{w}(f)}{\gamma(t,f)}\right) + \alpha_f(\mathbf{w}^{\mathrm{H}}(f)\mathbf{a}(f,\Omega_d) - 1), \tag{8}$$

where α_f is the Lagrangian multiplier. The estimate of the weight vector and the variances is now obtained by differentiating the cost function $\mathscr{C}(\boldsymbol{\beta}_f)$ with respect to the unknown parameters (while assuming the other unknown parameters are constant) and setting the differentiation to zero. So, differentiating $\mathscr{C}(\boldsymbol{\beta}_f)$ with respect to $\mathbf{w}(f)$ and setting it to zero results in

$$\frac{\partial\mathscr{C}(\boldsymbol{\beta}_f)}{\partial\mathbf{w}(f)} = \sum_{t=1}^{N_t}\left(\frac{\mathbf{x}(t,f)\mathbf{x}^{\mathrm{H}}(t,f)}{\gamma(t,f)}\right)\mathbf{w}(f) + \alpha_f\mathbf{a}(f,\Omega_d) = 0. \tag{9}$$

Solving (9) gives an estimate of the weight vector of the MLDR beamformer as

$$\mathbf{w}(f) = \frac{\tilde{\mathbf{R}}_{\mathbf{x}}^{-1}(f)\mathbf{a}(f,\Omega_d)}{\mathbf{a}^{\mathrm{H}}(f,\Omega_d)\tilde{\mathbf{R}}_{\mathbf{x}}^{-1}(f)\mathbf{a}(f,\Omega_d)}, \tag{10}$$

where $\tilde{\mathbf{R}}_{\mathbf{x}}(f) = \frac{1}{N_t}\sum_{t=1}^{N_t}\left(\frac{\mathbf{x}(t,f)\mathbf{x}^{\mathrm{H}}(t,f)}{\gamma(t,f)}\right)$ is the array covariance matrix weighted by the variance of the prior distribution. To obtain the variances, $\mathscr{C}(\boldsymbol{\beta}_f)$ is differentiated with respect to $\gamma(t,f)$ and set to zero as

$$\frac{\partial\mathscr{C}(\boldsymbol{\beta}_f)}{\partial\gamma(t,f)} = \frac{1}{\gamma(t,f)} - \frac{|z(t,f)|^2}{\gamma^2(t,f)} = 0. \tag{11}$$

Solving (11), we get the estimate of $\gamma(t,f)$ as

$$\gamma(t,f) = |z(t,f)|^2 = |\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f)|^2. \tag{12}$$

Thus, the weight vector of the MLDR beamformer can be obtained by iteratively solving (10) and (12) until convergence or a fixed number of iterations. In this paper, we refer to the MLDR beamformer with a CG prior distribution as the MLDR-CG beamformer.

3. PROPOSED ACOUSTIC BEAMFORMER

In this section, we will introduce a new acoustic beamformer. To motivate the formulation of a new acoustic beamformer, we first have a look at the STFT spectrogram of a speech signal presented in Figure 1. From the figure, the sparse nature of the speech coefficients in the STFT domain is clearly visible. A vast majority of the STFT coefficients have zero or close-to-zero values. Only a few time-frequency (TF) bins contain significant coefficients of the speech signal. With this in mind, the proposed acoustic beamformer is formulated in such a way as to encourage sparsity at the beamformer output. To achieve the sparsity objective, we model the real and imaginary parts of the output of the proposed beamformer as random variables with zero-mean Laplacian

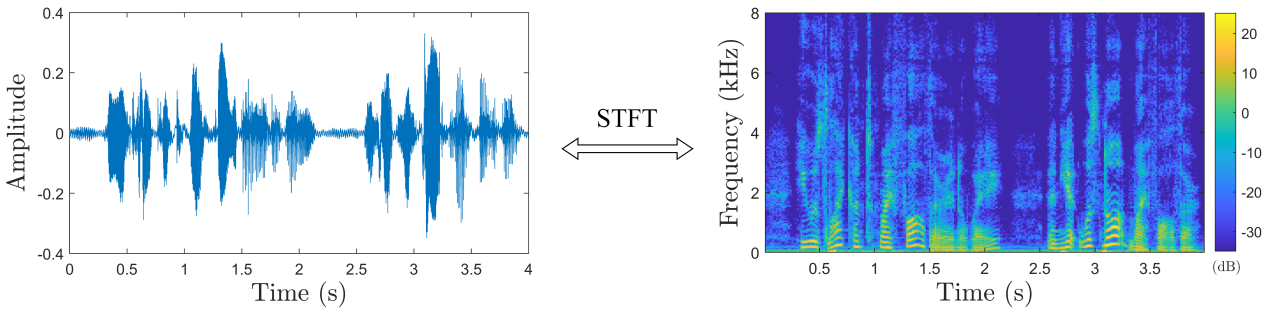

Figure 1: (left) Speech signal of a male speaker saying the words "*he was like unto my father in a way and yet was not my father*" from the Librispeech dataset [16] and (right) the spectrogram of the speech signal in the STFT domain. The sparse and non-stationarity nature of speech, the two features of speech utilised in our proposed beamformer, is clearly seen in the STFT domain. Any distribution that aims to model speech coefficients in the STFT domain should promote sparsity and have time-varying properties.

distributions. The Laplacian distribution has a heavier tail than the circular CG distribution and, as a result, promotes sparsity. This makes the Laplacian distribution more suitable to model speech coefficients than the CG distribution. Thus, the PDF of the output of the proposed beamformer is expressed as

$$\mathscr{P}(z(t,f)) = \frac{1}{\gamma(t,f)} e^{-2\frac{|\mathscr{R}(z(t,f))|+|\mathscr{I}(z(t,f))|}{\sqrt{\gamma(t,f)}}}, \tag{13}$$

where $\mathscr{R}(.)$ and $\mathscr{I}(.)$ denote the real and imaginary parts of a complex number, respectively. The PDF is designed in such a manner that the real and imaginary parts of the output of the proposed beamformer both follow a Laplacian distribution with a variance of $\sqrt{\gamma(t,f)}$. As was the case with MLDR-CG, we assume independence across time and frequency, which results in the joint PDF in a particular frequency bin to be

$$\mathscr{J}(\boldsymbol{\beta}_f) = \prod_{t=1}^{N_t} \mathscr{P}(z(t,f)). \tag{14}$$

Now, the negative of the log-likelihood of the joint PDF of the proposed beamformer is expressed as

$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta}_f) = -\log(\mathscr{P}(z(t,f))) &= \sum_{t=1}^{N_t} \left(\log(\gamma(t,f)) + 2\frac{|\mathscr{R}(z(t,f))| + |\mathscr{I}(z(t,f))|}{\sqrt{\gamma(t,f)}} \right) \\ &= \sum_{t=1}^{N_t} \left(\log(\gamma(t,f)) + 2\frac{|\mathscr{R}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))| + |\mathscr{I}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))|}{\sqrt{\gamma(t,f)}} \right).\end{aligned} \tag{15}$$

The cost function in (15) is not differentiable with respect to $\mathbf{w}(f)$. Hence, to get an estimate of the weight vector of the proposed beamformer, we form a linear programming problem. To do so, we collect all the terms in the cost function (15) containing the term $\mathbf{w}(f)$ as

$$\mathscr{W}(\mathbf{w}(f)) = \sum_{t=1}^{N_t} \frac{2}{\sqrt{\gamma(t,f)}} \left(|\mathscr{R}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))| + |\mathscr{I}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))| \right). \tag{16}$$

Then, the weight vector $\mathbf{w}(f)$ can be estimated by solving the following optimization problem

$$\min_{\mathbf{w}(f)} \quad \mathscr{W}(\mathbf{w}(f)) \qquad \text{s.t.} \qquad \mathbf{w}^{\mathrm{H}}(f)\mathbf{a}(f,\Omega_d) = 1. \tag{17}$$

Note that the constraint is the distortionless constraint towards the desired speaker. The optimization problem in (17) does not have a closed-form solution. Rather, the solution to (17) is

Algorithm 1 Proposed MLDR beamformer with Laplacian Prior (MLDR-LP)

inputs: $\mathbf{x}(t,f)\ \forall\ t$, Steering vector estimate $\mathbf{a}(f,\Omega_d)$, N_t, ϵ_f.
initialize: $\gamma(t,f) = |(\mathbf{w}^0(f))^{\mathrm{H}}\mathbf{x}(t,f)|^2$
repeat:
$\mathbf{w}(f) \leftarrow$ solve (18) with the estimated $\mathbf{a}(f,\Omega_d)$
$z(t,f) \leftarrow \mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f)$
$\gamma(t,f) \leftarrow \max\left\{\left(|\mathscr{R}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))| + |\mathscr{I}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))|\right)^2, \epsilon_f\right\}$
until: condition satisfied
output: beamformer weights $\mathbf{w}(f)$ for each f

obtained by solving the following linear programming problem

$$\begin{aligned}
\min_{\mathbf{h},\mathbf{w}(f)} \quad & \|\mathbf{h}\|_1 \\
\text{s.t.} \quad & \mathbf{h} \geq 0 \\
& |\mathscr{R}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))| \leq \frac{\sqrt{\gamma(t,f)}}{2} h_{2t-1} \\
& |\mathscr{I}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))| \leq \frac{\sqrt{\gamma(t,f)}}{2} h_{2t} \\
& \mathbf{w}^{\mathrm{H}}(f)\mathbf{a}(f,\Omega_d) = 1,
\end{aligned} \tag{18}$$

where $\mathbf{h} \in \mathbb{R}^{2N_t}$ is a dummy variable and h_t denotes the t^{th} element of the vector $\mathbf{h}$. (18) can be solved using any standard convex optimization toolkit, such as the CVX toolkit [17]. Now, since the cost function in (15) is differentiable with respect to $\gamma(t,f)$, the estimate of the variance of the Laplacian prior of the proposed beamformer is obtained by differentiating $\mathscr{L}(\beta_f)$ with respect to $\gamma(t,f)$ and setting it to zero as

$$\frac{\partial \mathscr{C}(\beta_f)}{\partial \gamma(t,f)} = \frac{1}{\gamma(t,f)} - \frac{|\mathscr{R}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))| + |\mathscr{I}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))|}{\gamma^{\frac{3}{2}}(t,f)} = 0. \tag{19}$$

This gives a closed-form solution for $\gamma(t,f)$ as

$$\gamma(t,f) = \left(|\mathscr{R}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))| + |\mathscr{I}(\mathbf{w}^{\mathrm{H}}(f)\mathbf{x}(t,f))|\right)^2. \tag{20}$$

Solving (18) and (20) iteratively until convergence or until a fixed number of iterations gives the weight vector of the proposed acoustic beamformer. Since the proposed beamformer is a maximum likelihood (ML) beamformer with a distortionless response (DR) and a Laplacian prior (LP), we refer to our beamformer as the MLDR-LP beamformer. The pseudocode of the proposed beamformer is presented in Algorithm 1. The weight vector at the zeroth iteration $\mathbf{w}^0(f)$ is initialized as an MPDR beamformer.

The appearance of the ℓ_1-norm in the linear program in (18) solidifies the assertion that the Laplacian prior promotes sparsity. This should result in the MLDR-LP being a better beamformer than the MLDR-CG beamformer. Promoting sparsity at the beamformer output is also important because the input signal will be dense in the STFT domain as it contains the speech from all the speakers as well as the microphone and background noise. However, after passing through the beamforming filter, the signal should become sparse as it should correspond to the speech of only the desired speaker.

4. PERFORMANCE EVALUATION

Let us consider a scenario where a circular microphone array of radius 10 cm with $P = 7$ microphones is placed at the centre of a room with dimensions $10 \times 10 \times 4$ m^3. Let us also consider

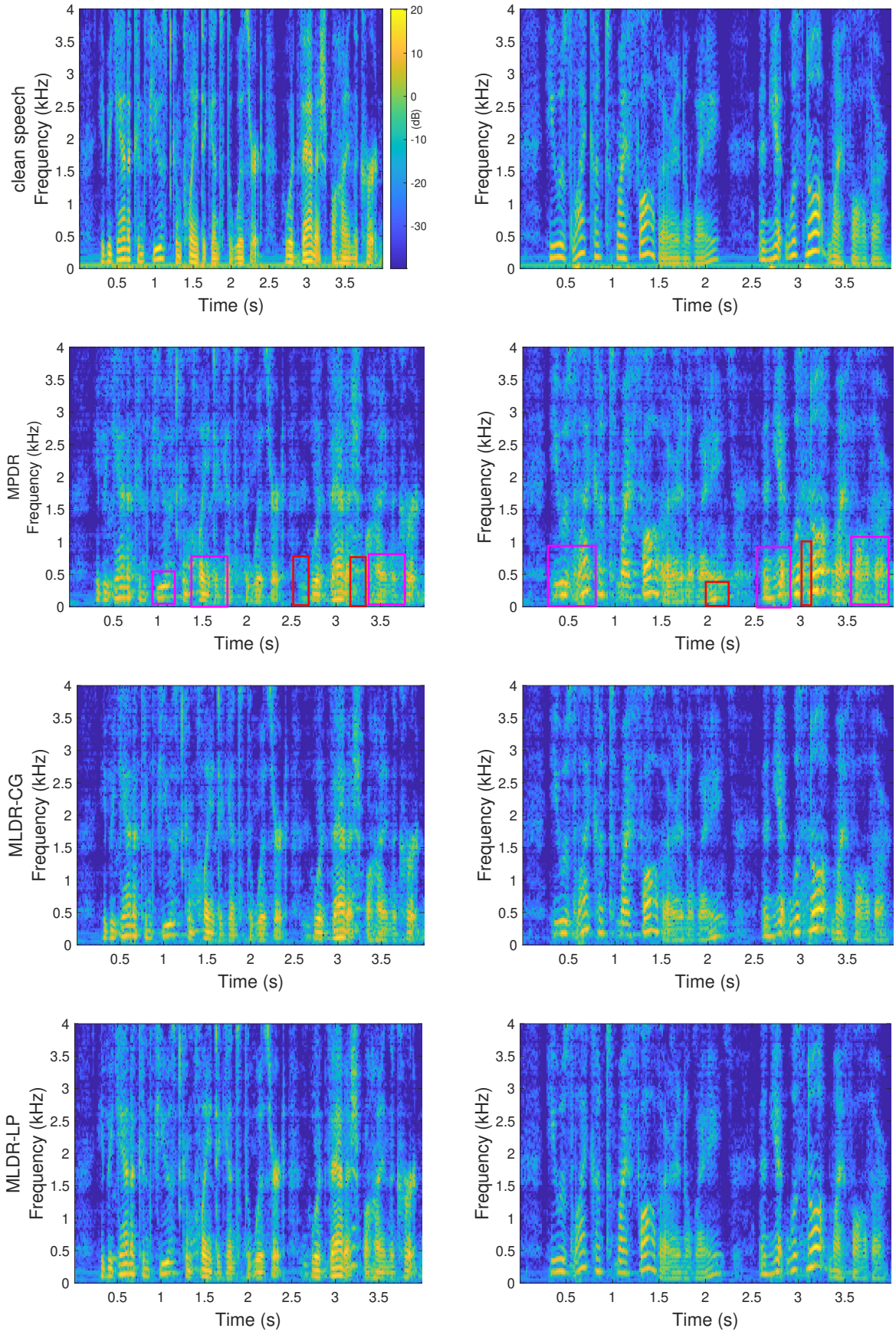

Figure 2: Spectrograms in the STFT domain of the clean speeches (first row) and the estimated speeches at the output of all the beamformers. The first and second columns correspond to the first and second speakers, respectively. The red boxes highlight the residuals of the interfering speech, and the pink boxes highlight the degradation of the speech coefficients.

Table 1: A table listing the objective measures comparing the speech at the output of all the beamformers with the clean speech signals in the time domain for different reverberation times and noiseless acoustic environment.

	Anechoic (AE)				RT_{60} = 200 ms			
Beamformers	**CD**	**PESQ**	**WSS**	**LLR**	**CD**	**PESQ**	**WSS**	**LLR**
MPDR	0.5828	1.6202	21.2036	0.0697	2.8087	1.5486	31.5663	0.3924
MLDR-CG	0.2995	2.2321	17.4104	0.0183	2.4299	2.1947	26.9594	0.2966
MLDR-LP	0.2513	2.5202	10.6512	0.0148	2.1753	2.2442	20.4928	0.2490

Table 2: A table listing the objective measures comparing the speech at the output of all the beamformers with the clean speech signals in the time domain for different reverberation times and Gaussian noise environment.

	Anechoic (AE)				RT_{60} = 200 ms			
Beamformers	**CD**	**PESQ**	**WSS**	**LLR**	**CD**	**PESQ**	**WSS**	**LLR**
MPDR	2.9672	1.5016	35.8892	0.4618	3.8856	1.4304	36.0045	0.5136
MLDR-CG	2.3687	2.1867	22.2443	0.3651	2.9690	2.0110	24.8645	0.3735
MLDR-LP	2.0317	2.2412	11.1178	0.3100	2.2758	2.1447	20.8906	0.3444

that $D = 2$ speakers are present at angular locations $(60°, 45°)$ and $(60°, 135°)$, respectively, at a distance of 3 m from the centre of the array. The clean speeches of the two speakers are taken from the Librispeech dataset [16]. The sampling frequency of the clean speeches is 16 kHz, and the duration is 4 s. The room impulse responses (RIRs) from the location of the speakers to microphones are generated using the image method [18, 19]. The microphone signals are generated by convolving the clean speeches with the respective RIRs. In the STFT domain, we kept the frame length to be of 512 samples with an overlap between consecutive frames of 50%. As a result, the total number of frames in each frequency bin was $N_t = 256$.

Figure 2 shows the STFT spectrograms of the clean speeches of both the speakers as well as the spectrograms at the output of the MPDR, MLDR-CG and the proposed MLDR-LP beamformers. The reverberation time (RT_{60}) for the RIRs was set to 200 ms. From the figure, we can see that the maximum likelihood beamformers, i.e. MLDR-CG and MLDR-LP, are better at filtering out the interfering speech residuals at the beamformer output than the MPDR beamformer. The MPDR beamformer also degrades the speech coefficients more than the maximum likelihood beamformers. Between MLDR-CG and MLDR-LP, we can see that our proposed beamformer is closer to the clean speech signals with fewer degradation and distortions. Thus, the proposed beamformer performs the best in preserving the desired speech signal and filtering out the interfering speech in a noiseless acoustic scenario.

Since the visual results in Figure 2 are not conclusive, we present objective measures comparing the estimated clean speech signal of speaker 1 (i.e. the speaker at location $(60°, 45°)$) at the output of all the beamformers with the clean speech signal of speaker 1. The objective measures computed include the Perceptual Evaluation of Speech Quality (PESQ), Cepstral Distance (CD), Weighted-Slope Spectral distance (WSS) and the Log-Likelihood Ratio (LLR) [20].

Table 3: A table listing the objective measures comparing the speech at the output of all the beamformers with the clean speech signals in the time domain for different reverberation times and Laplacian noise environment.

	Anechoic (AE)				RT_{60} = 200 ms			
Beamformers	**CD**	**PESQ**	**WSS**	**LLR**	**CD**	**PESQ**	**WSS**	**LLR**
MPDR	3.6851	1.1412	37.6732	0.4686	3.8959	1.4308	39.5663	0.5024
MLDR-CG	2.6819	1.8958	22.3719	0.3734	2.9660	2.0153	25.3170	0.3837
MLDR-LP	2.2117	2.2373	15.8156	0.3174	2.2902	2.1069	20.2211	0.3490

Table 1 presents the objective measures for the noiseless case. From the table, we can deduce that the MPDR beamformer performs the worst, whereas the proposed MLDR-LP beamformer performs the best in a noiseless acoustic scenario. This is because the proposed beamformer encourages sparsity, which results in better modelling of the desired clean speech signals. Also, the maximum likelihood based beamformers take into account the variance in each TF bin. This weighting of the cost function with the variance means that while minimizing, the TF bins with lower variances (that typically correspond to TF bins with no desired speech component) are given higher weightage than the TF bins with higher variance (that typically have desired speech components). This results in a better elimination of the interfering speech residuals. This feature, together with the promotion of sparsity, makes the proposed MLDR-LP beamformer perform the best among all the beamformers discussed in this work in a noiseless scenario.

In the precious simulations, no noise was added to the array signals. Table 2 and Table 3 present the objective measures of the estimated speech of speaker 1 at the output of all the beamformers compared to the clean speech signal in the presence of normalized Gaussian and Laplacian noise, respectively. The signal-to-noise ratio (SNR) was set to be 30 dB. The results are presented for the anechoic scenario as well as for an acoustic scenario with a reverberation time of 200 ms. From the two tables, we can see that even in the presence of both kinds of noise, the proposed beamformer outperforms the other beamformers.

5. CONCLUSION

Acoustic beamforming can be used to capture the speech of the desired speaker when interfering speakers are speaking simultaneously using an array of microphones. Usually, acoustic beamforming is performed by modelling the desired speech at the output of the beamformer as a circular complex Gaussian (CG) random variable. However, clean speech is sparse in the STFT domain, and the CG distribution does not give a sparse solution. In this work, we propose a new acoustic beamformer that promotes sparsity at the beamformer output. We propose to model the real and imaginary parts of the beamformer output as random variables that follow Laplacian distributions. To account for the non-stationarity of speech signals, the variance of the prior distribution is designed to be time and frequency varying. A distortionless constraint is also added to the formulation of the proposed beamformer. A maximum likelihood technique is used to obtain the weights of the proposed beamformer as well as the variances of the prior distribution. To obtain the weight vector, a linear program has to be solved while a closed-formed solution exists for the estimation of the variances. Simulation results show that the proposed beamformer outperforms the existing beamformer.

ACKNOWLEDGEMENTS

This work was supported by the Department of Science and Technology, Government of India under the MATRICS Scheme (MTR/2022/000290) and the TEOCO Chair of the Indian Institute of Technology Gandhinagar.

REFERENCES

1. Sharon Gannot, David Burshtein, and Ehud Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626, 2001.
2. Jacob Benesty, Jingdong Chen, and Yiteng Huang. *Microphone array signal processing*, volume 1. Springer Science & Business Media, 2008.
3. Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1717–1731, 2010.
4. Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Speech dereverberation in short time fourier transform domain with crossband effect compensation. In *2008 Hands-Free Speech Communication and Microphone Arrays*, pages 220–223. IEEE, 2008.
5. Shekhar Kumar Yadav and Nithin V George. Joint dereverberation and beamforming with blind estimation of the shape parameter of the desired source prior. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:779–793, 2023.
6. Tomohiro Nakatani and Keisuke Kinoshita. Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation. In *27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.
7. Tomohiro Nakatani, Christoph Boeddeker, Keisuke Kinoshita, Rintaro Ikeshita, Marc Delcroix, and Reinhold Haeb-Umbach. Jointly optimal denoising, dereverberation, and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2267–2282, 2020.
8. Byung Joon Cho, Jun-Min Lee, and Hyung-Min 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*, 26(9):1398–1402, 2019.
9. Byung Joon Cho and Hyung-Min Park. Convolutional maximum-likelihood distortionless response beamforming with steering vector estimation for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1352–1367, 2021.
10. Ante Jukić and Simon Doclo. Speech dereverberation using weighted prediction error with laplacian model of the desired signal. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5172–5176. IEEE, 2014.
11. Shekhar Kumar Yadav and Nithin V George. Speech enhancement via maximum likelihood modal beamforming with complex gaussian and laplacian priors. In *31st European Signal Processing Conference (EUSIPCO)*, pages 26–30. IEEE, 2023.
12. Ante Jukić, Toon van Waterschoot, Timo Gerkmann, and Simon Doclo. Multi-channel linear prediction-based speech dereverberation with sparse priors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1509–1520, 2015.
13. Shekhar Kumar Yadav and Nithin V George. Sparse distortionless modal beamforming for spherical microphone arrays. *IEEE Signal Processing Letters*, 29:2068–2072, 2022.
14. Marcin Witkowski and Konrad Kowalczyk. Split bregman approach to linear prediction based dereverberation with enforced speech sparsity. *IEEE Signal Processing Letters*, 28:942–946, 2021.

15. Shekhar Kumar Yadav and Nithin V George. Distortionless acoustic beamforming with enhanced sparsity based on reweighted ℓ_1-norm minimization. In *International Congress on Acoustics (ICA)*, pages A05–227, 2022.
16. Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210, 2015.
17. M. Grant and S. Boyd. CVX: MATLAB software for disciplined convex programming. http://cvxr.com/cvx/. [Access on: April 2024].
18. Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.
19. E.A.P. Habets. Room impulse response (RIR) generator. https://github.com/ehabets/RIR-Generator. [Access on: April 2024].
20. Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.