MAXIMUM-LIKELIHOOD SOUND SOURCE LOCALIZATION WITH A MULTIVARIATE COMPLEX LAPLACIAN DISTRIBUTION

Bowon Lee, Ton Kalker, and Ronald W. Schafer

Hewlett-Packard Laboratories
1501 Page Mill Road, Palo Alto, CA 94304, USA

ABSTRACT

Traditional Maximum Likelihood Sound Source Localization (ML-SSL) methods assume that the Fourier coefficients of signals have a Gaussian distribution. In many practical speech processing applications, the discrete Fourier transform (DFT) coefficients are computed from finite duration signals, which makes the Gaussian assumption less favorable choice for signals whose time-domain distributions are non-Gaussian. Recently, for audio signals including speech, distributions such as Laplacian or Gamma distribution have been shown to better model the time-domain samples and their DFT coefficients. Motivated by this, we propose a new ML-SSL method based on a multivariate complex Laplacian distribution.

Index Terms— Sound Source Localization, Maximum-Likelihood Estimation, Multivariate Laplacian Distribution

1. INTRODUCTION

Sound source localization (SSL) is an important topic for hands-free speech communication systems using a microphone array [1]. For a pair of microphones, time delay of arrival (TDOA) can be computed using the generalized cross correlation (GCC) method with the phase transform (PHAT) or maximum likelihood (ML) prefilters [2]. For SSL with more than two microphones, we can use the steered response power (SRP) method as an extension of the GCC [3], TDOA based projection [4], or the ML estimation by modeling the multivariate distribution of signals [5].

Motivated by the central limit theorem as well as mathematical tractability, maximum likelihood sound source localization (ML-SSL) assumes the Gaussian distribution for the discrete Fourier coefficients of signals [2, 5]. In practice however, a Gaussian model is not always the most accurate description of discrete Fourier coefficients. In particular, it has been reported that the distribution of time-domain speech samples is well represented by Laplacian distribution [6]. It has been also reported that the frequency components of speech samples are better modeled by distributions such as Laplacian [7], Gamma [8], or generalized Gaussian [9] than the Gaussian distribution in the context of speech enhancement, voice activity detection, and blind source separation, all of which motivate use of non-Gaussian distributions for the ML-SSL.

Compared to the aforementioned scenarios requiring univariate real or complex distributions, ML-SSL requires a multivariate complex distribution, which cannot be uniquely defined for non-Gaussian cases. Eltoft et al. [10] proposed a multivariate Laplacian distribution as a multivariate scale mixture of Gaussians using an exponential scale factor.

In this paper, we propose a multivariate complex Laplacian distribution based on [10], and then a new ML-SSL method using this distribution. We compare our proposed method to the ML-SSL based on the likelihood function with multivariate complex Gaussian distribution assumption.

2. BACKGROUND

For an array of M microphones with source signal $s(t)$, signal $x_m(t)$ captured at the m^{th} microphone can be expressed as

$$x_m(t) = h_m(t) * s(t) + n_m(t), \quad m = 1, \cdots, M \tag{1}$$

where $h_m(t)$ and $n_m(t)$ denote impulse response and noise at the m^{th} microphone and $*$ denotes convolution. We decompose the impulse response as $h_m(t) = d_m(t) + r_m(t)$ with $d_m(t)$ and $r_m(t)$ representing delay and reverberation respectively. We can express received signals as a vector in the frequency domain

$$\mathbf{X}_f = S(f)\mathbf{D}_f + S(f)\mathbf{R}_f + \mathbf{N}_f \tag{2}$$

where

$$\mathbf{X}_f = [X_1(f), X_2(f), \cdots, X_M(f)]^T$$
$$\mathbf{R}_f = [R_1(f), R_2(f), \cdots, R_M(f)]^T$$
$$\mathbf{N}_f = [N_1(f), N_2(f), \cdots, N_M(f)]^T$$

with each element denoting the discrete Fourier transform of the corresponding signals and

$$\mathbf{D}_f = [\alpha_1(f)e^{-j2\pi f\tau_1}, \alpha_2(f)e^{-j2\pi f\tau_2}, \cdots, \alpha_M(f)e^{-j2\pi f\tau_M}]^T \tag{3}$$

denoting the delay vector as point-wise multiplication of attenuation $\alpha_m(f)$ and time delay τ_m for $m = 1, 2, \cdots, M$.

2.1. Maximum-Likelihood Sound Source Localization

With the signal model described above, the ML-SSL problem for a single frequency f is to find a delay vector $\mathbf{D}_f$ that maximizes the likelihood of observing signal vector $\mathbf{X}_f$, i.e.,

$$\hat{\mathbf{D}}_f = \arg\max_{\mathbf{D}_f} \mathcal{L}(\mathbf{X}_f|\mathbf{D}_f) \tag{4}$$

where $\mathcal{L}(\mathbf{X}_f|\mathbf{D}_f) \propto \log p(\mathbf{X}_f|\mathbf{D}_f)$ denotes a log-likelihood function of observing $\mathbf{X}_f$ given $\mathbf{D}_f$ with respect to a properly modeled $p(\mathbf{X}_f|\mathbf{D}_f)$. In the case of using multiple frequency components, adopting the common assumption that the probability distribution is independent among different frequency components, i.e., the pdf can be expressed as a product of those of individual frequency components, then we have

$$\hat{\mathrm{l}} = \arg\max_{\mathcal{D}_\mathrm{l}} \sum_{\mathbf{D}_f \in \mathcal{D}_\mathrm{l}} \mathcal{L}(\mathbf{X}_f|\mathbf{D}_f) \tag{5}$$

where $\mathcal{D}_\mathrm{l}$ denotes a set of delay vectors $\mathbf{D}_f$ corresponding to the source location l.

2.2. ML-SSL with the Gaussian distribution

In order to formulate the joint pdf of signals $\mathbf{X}_f$, we consider the source speech $S(f)$ as deterministic, given signal $\mathbf{X}_f$ and the delay $\mathbf{D}_f$, whereas reverberation $\mathbf{R}_f$ and background noise $\mathbf{N}_f$ are stochastic, both with zero mean. If we use a Gaussian assumption for the stochastic parts, then the pdf of $\mathbf{X}_f$ given $\mathbf{D}_f$ can be expressed as [5]

$$p(\mathbf{X}_f|\mathbf{D}_f) \propto \exp\left\{-\frac{1}{2}[\mathbf{X}_f - S(f)\mathbf{D}_f]^H \mathbf{Q}_f^{-1}[\mathbf{X}_f - S(f)\mathbf{D}_f]\right\} \tag{6}$$

i.e., a complex Gaussian with mean $S(f)\mathbf{D}_f$ and covariance matrix $\mathbf{Q}_f$ defined as

$$\begin{aligned}\mathbf{Q}_f &= E\left\{[\mathbf{X}_f - S(f)\mathbf{D}_f][\mathbf{X}_f - S(f)\mathbf{D}_f]^H\right\} \\ &= E\{\mathbf{X}_f\mathbf{X}_f^H\} - |S(f)|^2\mathbf{D}_f\mathbf{D}_f^H\end{aligned} \tag{7}$$

where ^{H} denotes Hermitian transpose. Provided that the covariance matrix $\mathbf{Q}_f$ is available, we now need to estimate $S(f)$ given $\mathbf{X}_f$ and $\mathbf{D}_f$. Zhang et al. [5] derived an estimate of $S(f)$ which maximizes the Gaussian probability in Eq. (6)

$$\hat{S}(f) = \frac{\mathbf{D}_f^H\mathbf{Q}_f^{-1}\mathbf{X}_f}{\mathbf{D}_f^H\mathbf{Q}_f^{-1}\mathbf{D}_f}. \tag{8}$$

If we take the log-likelihood of Eq. (6) and use $\hat{S}(f)\mathbf{D}_f$ as its mean, we have

$$\mathcal{L}_G(\mathbf{X}_f|\mathbf{D}_f) = -[\mathbf{X}_f - \hat{S}(f)\mathbf{D}_f]^H\mathbf{Q}_f^{-1}[\mathbf{X}_f - \hat{S}(f)\mathbf{D}_f] \tag{9}$$

and the Maximum Likelihood solution for the Gaussian distribution has been shown to be [5]

$$\hat{\mathrm{l}} = \arg\max_{\mathcal{D}_\mathrm{l}} \sum_{\mathbf{D}_f \in \mathcal{D}_\mathrm{l}} \frac{[\mathbf{D}_f^H\mathbf{Q}_f^{-1}\mathbf{X}_f]^H\mathbf{D}_f^H\mathbf{Q}_f^{-1}\mathbf{X}_f}{\mathbf{D}_f^H\mathbf{Q}_f^{-1}\mathbf{D}_f}. \tag{10}$$

2.3. Covariance matrix estimation

In order to find the ML-SSL solution, we need to estimate the covariance matrix $\mathbf{Q}_f$. According to Eqs. (2) and (7) with an assumption that $\mathbf{R}_f$ and $\mathbf{N}_f$ are uncorrelated, we find [5]

$$\mathbf{Q}_f = |S(f)|^2 E\{\mathbf{R}_f\mathbf{R}_f^H\} + E\{\mathbf{N}_f\mathbf{N}_f^H\}. \tag{11}$$

Provided that we have $E\{\mathbf{N}_f\mathbf{N}_f^H\}$ by estimating it from available noise-only data, we approximate $|S(f)|^2E\{\mathbf{R}_f\mathbf{R}_f^H\}$ as a fraction of the difference between $E\{\mathbf{X}_f\mathbf{X}_f^H\}$ and $E\{\mathbf{N}_f\mathbf{N}_f^H\}$ [5]

$$|S(f)|^2 E\{\mathbf{R}_f\mathbf{R}_f^H\} \approx \lambda\left(E\{\mathbf{X}_f\mathbf{X}_f^H\} - E\{\mathbf{N}_f\mathbf{N}_f^H\}\right) \tag{12}$$

for $0 < \lambda < 1$ and use $E\{\mathbf{X}_f\mathbf{X}_f^H\} = \mathbf{X}_f\mathbf{X}_f^H$. We can also use the diagonal covariance matrix assumption [5]

$$\hat{\mathbf{Q}}_f = \mathrm{diag}\left(q_1(f), q_2(f), \cdots, q_M(f)\right) \tag{13}$$

where

$$q_m(f) = \lambda|X_m(f)|^2 + (1-\lambda)E\{|N_m(f)|^2\}. \tag{14}$$

3. PROPOSED METHOD

The joint complex Gaussian distribution in Eq. (6) assumes uniformly distributed phase, which gives a closed-form expression for the multivariate complex Gaussian distribution in terms of a vector of complex signals $\mathbf{X}_f$.Moreover, the commonly adopted assumption of uncorrelatedness, i.e., diagonal covariance essentially makes the multivariate distribution a product of independent pdfs. However, non-Gaussian multivariate distributions cannot be uniquely defined and their uncorrelatedness does not guarantee independence. Using a multivariate Laplacian pdf derived as a scaled mixture of Gaussian proposed in [10], we derive a closed-form expression of multivariate complex Laplacian pdf and propose an ML-SSL method based on the proposed distribution.

3.1. Multivariate Complex Laplacian distribution

For an $M \times 1$ random vector $\mathbf{Y}$, Eltoft et al. [10] proposed a multivariate Laplacian as a scaled mixture of Gaussian such that

$$\mathbf{Y} = \mathbf{y}_\mu + \sqrt{Z}\Gamma^{\frac{1}{2}}\mathbf{V} \tag{15}$$

where $\mathbf{V}$ is a $M \times 1$ zero mean Gaussian random vector with an identity covariance matrix, Z is a exponential random variable with mean $2/\sigma^2$, and Γ is a positive definite matrix with unity determinant and interpreted as an internal covariance structure of $\mathbf{Y}$. From Eq. (15), they derived a multivariate Laplacian pdf as

$$p(\mathbf{y}) = \frac{\sigma^2}{(2\pi)^{(M/2)}} \frac{K_{(M/2)-1}\left(\sigma||\mathbf{y} - \mathbf{y}_\mu||_\Gamma\right)}{\left(\frac{1}{\sigma}||\mathbf{y} - \mathbf{y}_\mu||_\Gamma\right)^{(M/2)-1}} \tag{16}$$

where $K_b(\mathbf{y})$ denotes the modified Bessel function of the second kind with order b and $||\mathbf{y}-\mathbf{y}_\mu||_\Gamma$ is the Mahalanobis distance defined as

$$||\mathbf{y}-\mathbf{y}_\mu||_\Gamma = \sqrt{(\mathbf{y}-\mathbf{y}_\mu)^T\Gamma^{-1}(\mathbf{y}-\mathbf{y}_\mu)}. \quad (17)$$

For an $M\times 1$ multivariate complex variable $\mathbf{X}_f$ with mean $\bar{\mathbf{X}}_f$ and an $M\times M$ positive definite Hermitian matrix $\mathbf{C}_f$, we can define the Mahalanobis distance as

$$||\mathbf{X}_f-\bar{\mathbf{X}}_f||_{\mathbf{C}_f} = \sqrt{(\mathbf{X}_f-\bar{\mathbf{X}}_f)^H\mathbf{C}_f^{-1}(\mathbf{X}_f-\bar{\mathbf{X}}_f)}. \quad (18)$$

Since $(\mathbf{X}_f-\bar{\mathbf{X}}_f)^H\mathbf{C}_f^{-1}(\mathbf{X}_f-\bar{\mathbf{X}}_f)$ in Eq. (18) is always real and positive and a quadratic formula for M complex variables, it is equivalent to a quadratic formula for $2M$ real variables with a $2M\times 2M$ real covariance matrix. Therefore, the complex Mahalanobis distance of a $M\times 1$ complex vector in Eq. (18) is equivalent to a real Mahalanobis distance of $2M\times 1$ real vector with a corresponding $2M\times 2M$ real covariance matrix. Hence, we can express a multivariate complex Laplacian pdf of $\mathbf{X}_f$ by Eq. (17) for Eq. (18) and replacing M with $2M$ in Eq. (16) as

$$p(\mathbf{X}_f) = \frac{\sigma^2}{(2\pi)^M}\frac{K_{M-1}\left(\sigma||\mathbf{X}_f-\bar{\mathbf{X}}_f||_{\mathbf{C}_f}\right)}{\left(\frac{1}{\sigma}||\mathbf{X}_f-\bar{\mathbf{X}}_f||_{\mathbf{C}_f}\right)^{M-1}}. \quad (19)$$

3.2. ML-SSL with a complex Laplacian distribution

We can define a log-likelihood function based on the Laplacian distribution in Eq. (19) as

$$\begin{aligned}\mathcal{L}_L(\mathbf{X}_f|\mathbf{D}_f) = &\log\left\{K_{M-1}\left(\sigma||\mathbf{X}_f-\bar{\mathbf{X}}_f||_{\mathbf{C}_f}\right)\right\}\\ &-(M-1)\log\left(||\mathbf{X}_f-\bar{\mathbf{X}}_f||_{\mathbf{C}_f}\right).\end{aligned} \quad (20)$$

with mean $\bar{\mathbf{X}} = S(f)\mathbf{D}_f$. Since the likelihood is conditioned upon the delay vector $\mathbf{D}_f$, we can use the minimum variance distortionless response (MVDR) of speech $S(f)$ for the direction corresponding to $\mathbf{D}_f$ to estimate $\bar{\mathbf{X}}$.

Suppose that we have an optimal beamformer $\bar{\mathbf{W}}_f$ to reconstruct $S(f)$

$$\hat{S}(f) = \bar{\mathbf{W}}_f^H\mathbf{X}_f \quad (21)$$

in the sense that the reconstructed source speech $\hat{S}(f)$ is distortionless response corresponding to $\mathbf{D}_f$, with a constraint $\mathbf{W}_f^H\mathbf{D}_f = 1$ while maximizing the signal-to-noise ratio (SNR) by minimizing the overall variance such that

$$\begin{aligned}\bar{\mathbf{W}}_f &= \arg\min_{\mathbf{W}_f} E\{|\mathbf{W}_f^H\mathbf{X}_f|^2\}\\ &= \arg\min_{\mathbf{W}_f} \mathbf{W}_f^H E\{\mathbf{X}_f\mathbf{X}_f^H\}\mathbf{W}_f\\ &= \arg\min_{\mathbf{W}_f}\left[\mathbf{W}_f^H\mathbf{Q}_f\mathbf{W}_f + |S(f)|^2|\mathbf{W}_f^H\mathbf{D}_f|^2\right]\\ &= \arg\min_{\mathbf{W}_f}\mathbf{W}_f^H\mathbf{Q}_f\mathbf{W}_f.\end{aligned} \quad (22)$$

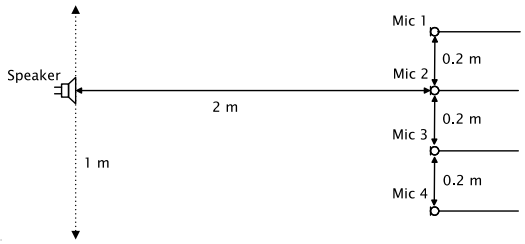

Fig. 1. Experimental setup showing source location relative to microphone array.

It has been shown that the solution for Eq. (22) can be expressed as [1]

$$\bar{\mathbf{W}}_f = \frac{\mathbf{Q}_f^{-1}\mathbf{D}_f}{\mathbf{D}_f^H\mathbf{Q}_f^{-1}\mathbf{D}_f} \quad (23)$$

which gives

$$\hat{S}(f) = \frac{\mathbf{D}_f^H\mathbf{Q}_f^{-1}\mathbf{X}_f}{\mathbf{D}_f^H\mathbf{Q}_f^{-1}\mathbf{D}_f} \quad (24)$$

which is equivalent to the ML estimation of $S(f)$ with a Gaussian pdf in Eq. (8).

For matrix $\mathbf{C}_f$ acting as a mapping function from multivariate Gaussian to Laplacian in Eq. (15), we estimate it by normalizing the covariance matrix $\mathbf{Q}_f$ to have unity determinant such that

$$\mathbf{C}_f = \frac{\mathbf{Q}_f}{|\mathbf{Q}_f|^{1/M}}. \quad (25)$$

Therefore, the ML-SSL estimate with multivariate complex Laplacian distribution can be found as follows

$$\begin{aligned}\hat{\mathbf{l}} = \arg\max_{\mathcal{D}_\mathbf{l}} \sum_{\mathbf{D}_f\in\mathcal{D}_\mathbf{l}} \Big[&\log\left\{K_{M-1}\left(\sigma||\mathbf{X}_f-\bar{\mathbf{X}}_f||_{\mathbf{C}_f}\right)\right\}\\ &-(M-1)\log\left(||\mathbf{X}_f-\bar{\mathbf{X}}_f||_{\mathbf{C}_f}\right)\Big]\end{aligned} \quad (26)$$

with $\bar{\mathbf{X}}_f = \hat{S}(f)\mathbf{D}_f$ from Eq. (24) and $\mathbf{C}_f$ from Eq. (25).

4. EXPERIMENTS

In order to demonstrate our proposed method, we made a clean speech recording of a female speaker at 48 kHz sampling rate for ten seconds. Then we played it through a loudspeaker in a reverberant room and recorded signals with a four microphone uniform linear array having 0.2 m inter-microphone distance and located 2 m away from the loudspeaker as depicted in Fig. 1. We then degraded the captured signals with additive white Gaussian noise by varying the SNR from 6 dB to 24 dB in 6 dB increments. For source localization, we chose four non-overlapping windows with duration of 25 ms, 50 ms, 75 ms, and 100 ms and ran ML-SSL with the likelihood function based on the Gaussian assumption in Eq. (10) and the proposed Laplacian assumption in Eq. (26) with $\sigma = 4$. We ran two sets of experiments, one

Window size	25 ms		50 ms		75 ms		100 ms	
SNR	Gaussian	Laplacian	Gaussian	Laplacian	Gaussian	Laplacian	Gaussian	Laplacian
6	37.34	36.59	44.22	49.25	55.30	59.85	56.57	65.66
12	48.37	59.15	55.28	72.36	62.88	78.79	64.65	85.86
18	49.87	74.44	56.28	85.43	66.67	93.18	68.69	96.97
24	52.63	87.47	57.29	93.97	68.94	99.24	68.69	100.00
Overall	47.06	64.41	53.27	75.25	63.45	82.77	64.65	87.12

Table 1. Experimental results in % accuracy for ML-SSL with identity covariance matrices.

Window size	25 ms		50 ms		75 ms		100 ms	
SNR	Gaussian	Laplacian	Gaussian	Laplacian	Gaussian	Laplacian	Gaussian	Laplacian
6	30.58	36.09	37.69	47.74	37.88	52.27	47.47	62.63
12	45.61	54.14	56.78	67.34	66.67	74.24	73.74	84.85
18	66.92	69.67	80.40	84.92	86.36	91.67	92.93	97.98
24	84.71	84.46	92.96	93.47	97.73	97.73	100.00	100.00
Overall	56.95	61.09	66.96	73.37	72.16	78.98	78.54	86.36

Table 2. Experimental results in % accuracy for ML-SSL with estimated covariance matrices

with an identity matrix $\mathbf{I}$ for $\mathbf{Q}_f$ and $\mathbf{C}_f$ and the other with $\mathbf{Q}_f$ estimated using Eq. (13) with $\lambda = 0.2$ and $\mathbf{C}_f$ with Eq. (25). Performance was evaluated for each frame by computing likelihood at each of 26 evenly spaced points along the dotted line in Fig. 1. The estimate was considered correct if the ML estimate occured at the actual location. The % accuracy over all frames is summerized in Tables 1 and 2.

We observe that the Laplacian model performs consistently better than the Gaussian model. In the Gaussian case we find that for 6 dB SNR, setting $\mathbf{Q}_f = \mathbf{I}$ gives better result than using the estimate of Eq. (13). In the Laplacian case we find that $\mathbf{C}_f = \mathbf{I}$ performs better than the estimate of Eq. (25) across all SNRs.

5. CONCLUSION

In this paper, we proposed a multivariate complex Laplacian pdf for the ML-SSL and demonstrated that it outperforms the ML-SSL with the Gaussian distribution assumption. We also discovered that for low SNR, an identity covariance matrix gives better performance than its estimation, which indicates that its estimate becomes less accurate for low SNRs. It is important to note that the internal covariance matrix estimation for the proposed multivariate complex Laplacian pdf is crucial for its performance and needs to be further investigated.

6. REFERENCES

[1] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, Germany, 2001.

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time-delay," *IEEE Trans. Acoust., Speech and Audio Process.*, vol. ASSP-24, no. 4, pp. 320–327, 1976.

[3] J. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments*, Ph.D. thesis, Brown University, Providence, RI, May 2000.

[4] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech and Audio Process.*, vol. 5, pp. 45–50, 1997.

[5] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, 2007, vol. I, pp. 125–128.

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

[7] J.-H. Chang and N. S. Kim, "Voice activity detection based on complex Laplacian model," *Electron. Letters*, vol. 39, no. 7, pp. 632–634, 2003.

[8] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, vol. 1, pp. 253–256, 2002.

[9] J.-H. Chang, J. W. Shin, and N. S. Kim, "Voice activity detector employing generalised Gaussian distribution," *Electron. Letters*, vol. 40, no. 24, pp. 1561–1563, 2004.

[10] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate Laplace distribution," *IEEE Signal Process. Letters*, vol. 13, no. 5, pp. 300–303, 2006.