End-Fire Degradation–Robust DOA Estimation for Compact Linear Microphone Arrays

Zheng Wen[1], *Huayang Wang*[1], *Xin Guo*[1], *Gongping Huang*[1]

[1] Electronic Information School, Wuhan University, Wuhan, China

3191479712@qq.com, 2540160476@qq.com, 1919304068@qq.com,gongpinghuang@gmail.com

Abstract

Source Localization using compact linear microphone arrays exhibits a pronounced performance degradation when the source is located near the end-fire directions. This paper firstly provides a systematic analysis of this phenomenon and shows that the severe error increase near end-fire originates from the ill-conditioned nonlinear mapping between time-difference-of-arrival (TDOA) estimation and azimuth angles under the far-field plane-wave model, where small delay perturbations are strongly amplified into large angular errors. To quantitatively evaluate this effect in realistic acoustic conditions, a real-world dataset is collected using a compact linear microphone array deployed in a reverberant indoor environment. Several widely used Direction-of-arrival (DOA) estimation methods are evaluated on this dataset, revealing consistent and significant accuracy degradation near end-fire directions. Based on the above analysis and observations, an improved DOA estimation strategy is further proposed to mitigate end-fire degradation. The proposed method combines reliability-aware PHAT-based processing with physically constrained delay refinement and weighted fusion, effectively reducing error amplification near end-fire directions.

Index Terms: Source Localization,compact linear microphone arrays, end-fire degradation,reliability-aware PHAT-based processing, weighted fusion

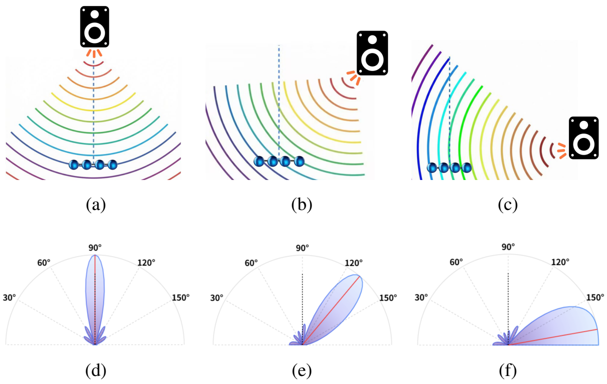
1. Introduction

Source Localization or Direction-of-arrival (DOA) estimation using microphone arrays is a fundamental task in many acoustic signal processing applications, including speech capture, human–machine interaction, and spatial audio analysis [1-4]. Among various array geometries, uniform linear arrays (ULAs) are particularly attractive and widely integrated into consumer devices such as televisions and home control terminals, where reliable far-field sound source localization is generally required. However, compact ULAs suffer from severe performance degradation when sound sources are located near the end-fire directions of the array. In practical deployments, it is frequently observed that localization accuracy near the array axis is significantly lower than that in side-fire directions. This phenomenon becomes especially pronounced for compact arrays with small apertures operating in indoor environments, where reverberation and background noise further distort inter-channel phase information [5]. Although end-fire degradation has been reported and analyzed in prior studies (e.g., end-fire MVDR analysis and end-fire DOA bias characterization in [6, 7]), it is often treated as an empirical artifact or a secondary limitation, rather than being systematically analyzed from a modeling and inversion perspective.

This paper firstly revisits the end-fire degradation problem from a theoretical standpoint. Under the commonly adopted far-field plane-wave model, DOA estimation with ULAs relies on inverting the nonlinear relationship between time-difference-of-arrival (TDOA) and azimuth angle. It is shown that this inversion becomes severely ill-conditioned near end-fire directions, where the sensitivity of the estimated angle to small TDOA perturbations increases dramatically. As a consequence, even minor estimation errors caused by noise, reverberation, or discretization can be amplified into large angular deviations. This analysis provides a clear explanation of why conventional DOA estimators, including SRP-PHAT [8,9] and its SRP-PHAT variants [10], MVDR beamforming [11], and MUSIC [12] exhibit unstable and biased behavior near end-fire, especially for compact arrays.

To evaluate this phenomenon under realistic acoustic conditions, this work further addresses a practical gap in existing studies: the lack of publicly available real-world datasets specifically designed to assess end-fire robustness. A dataset is therefore collected using a compact linear microphone array deployed in a reverberant indoor environment representative of typical home scenarios. Several widely used DOA estimation methods are systematically evaluated on this dataset, revealing consistent and significant error growth near end-fire directions across different algorithms and source distances. To promote reproducibility and facilitate fair comparison, the recorded dataset and evaluation methods are made publicly available.

Then, an improved DOA estimation strategy is proposed to mitigate end-fire degradation without increasing array aperture or relaxing the far-field assumption. The proposed approach directly targets the ill-conditioned nature of end-fire DOA inversion. By combining reliability-aware PHAT-based processing, physically constrained and oversampled TDOA estimation, and weighted fusion in the cosine domain prior to angle inversion, the proposed method effectively suppresses error amplification near end-fire directions. Experimental results on the real-recorded dataset demonstrate improved localization accuracy and stability in the challenging end-fire regions, while maintaining competitive performance over the full azimuth range. The main contributions of this work are summarized as follows: 1) A theoretical analysis that explains end-fire degradation in compact ULAs as an ill-conditioned TDOA-to-DOA inversion problem under the far-field model; 2) A publicly available real-world dataset and benchmark evaluation that reveal the severity of end-fire errors for commonly used DOA estimation methods; 3) An improved DOA estimation method that mitigates end-fire degradation through reliability-aware processing and stabilized inversion.



(a) (b) (c)

(d) (e) (f)

Figure 1: *Visualization diagram of beamforming.Top row: sound source (black speaker) and array (blue dots). Bottom row: Corresponding SRP energy spatial distribution. Columns represent different source angles. Note the broadening mainlobe and reduced curvature as the source approaches end-fire, indicating high sensitivity to perturbations.*

2. Signal Model and Problem Formulation

2.1. Far-field observation model

Consider a compact uniform linear array (ULA) with M microphones at known positions $\{\mathbf{r}_m\}_{m=1}^{M} \subset \mathbb{R}^2$ on the horizontal plane. We adopt a 2-D azimuth model with propagation direction $\mathbf{d}(\theta) = [\cos\theta,\ \sin\theta]^T$. Under a far-field single-source assumption [13], the short-time Fourier transform (STFT) at microphone m is

$$Y_m(f,\ell) = S(f,\ell)\, e^{-j2\pi f\, \tau_m(\theta_s)} + V_m(f,\ell), \tag{1}$$

where $f > 0$ is frequency, ℓ is the frame index, $S(f,\ell)$ is the source STFT at a reference point, and $V_m(f,\ell)$ summarizes additive noise. The far-field delay is $\tau_m(\theta) = \mathbf{r}_m^T \mathbf{d}(\theta)/c$, with speed of sound c. For a microphone pair $p = (m_1, m_2) \in \mathcal{P}$, define $\Delta\mathbf{r}_p = \mathbf{r}_{m_1} - \mathbf{r}_{m_2}$ and baseline $d_p = \|\Delta\mathbf{r}_p\|$.The far-field TDOA is

$$\tau_p(\theta) = \frac{\Delta\mathbf{r}_p^T \mathbf{d}(\theta)}{c}. \tag{2}$$

In speech applications, most energy typically lies within the conventional wideband range (roughly 80–8000 Hz at 16 kHz sampling). In this work, we estimate DOA over a band $\mathcal{B} = [f_{\min}, f_{\max}]$ with $f_{\min} = 800$ Hz and $f_{\max} = 4500$ Hz, which captures the dominant speech energy while avoiding very low-frequency instability. For compact arrays, frequencies above $c/(2s)$ (with adjacent inter-element spacing s) may introduce spatial aliasing; unless otherwise stated, we focus on $\mathcal{B}$ such that $f_{\max} < c/(2s)$ to isolate the dominant failure modes studied in this work.

2.2. End-fire degradation as an ill-conditioned inversion

Assume the ULA is aligned with the x-axis and measure $\theta \in [0, \pi]$ from the $+x$ direction (end-fire: $\theta \approx 0^\circ$ or 180°; side-fire: $\theta \approx 90^\circ$). Then $\Delta\mathbf{r}_p = [\Delta x_p, 0]^T$ and $|\Delta x_p| = d_p$, so (2) reduces to

$$\tau_p(\theta) = \frac{d_p}{c}\cos\theta. \tag{3}$$

The mapping $\theta = \arccos(c\tau_p/d_p)$ implies the sensitivity

$$\left|\frac{\partial\theta}{\partial\tau_p}\right| = \frac{c}{d_p|\sin\theta|}, \tag{4}$$

showing that as $\theta \to 0^\circ$ or 180° the same TDOA perturbation induces much larger angular error.

To quantify this limitation, we adopt a local equivalent model (used for local error analysis) $\widehat{\tau}_p = \tau_p(\theta) + \varepsilon_p$, with $\varepsilon_p \sim \mathcal{N}(0, \sigma_{\tau,p}^2)$, where $\widehat{\tau}_p$ is any locally consistent TDOA-related measurement (e.g., GCC-PHAT peak [14], phase residual, or SRP pairwise statistic) and $\sigma_{\tau,p}^2$ aggregates effective uncertainty due to noise, reverberation, and residual model mismatch. The Fisher information is

$$\mathcal{I}_p(\theta) = \frac{1}{\sigma_{\tau,p}^2}\left(\frac{\partial\tau_p(\theta)}{\partial\theta}\right)^2 = \frac{1}{\sigma_{\tau,p}^2}\left(\frac{d_p}{c}\sin\theta\right)^2, \tag{5}$$

yielding the CRLB [15]

$$\operatorname{var}(\widehat{\theta}) \geq \frac{1}{\mathcal{I}_p(\theta)} = \frac{c^2\sigma_{\tau,p}^2}{d_p^2\sin^2\theta}. \tag{6}$$

If multiple microphone pairs provide approximately independent local measurements, the total Fisher information adds across pairs.

Thus, $\sin\theta \to 0$ implies $\mathcal{I}_p(\theta) \to 0$ and the CRLB diverges, explaining the intrinsic end-fire difficulty. Moreover, the nonlinear transform $\theta = \arccos(u)$ ($u = c\tau_p/d_p$) implies that even if u is approximately Gaussian, the induced $p_\theta(\theta) = p_u(\cos\theta)\sin\theta$ becomes *qualitatively* skewed/heavy-tailed near end-fire, consistent with the flattened peaks in Fig. 1.

3. Proposed Endfire-Robust PHAT Method

A broad class of steered response power (SRP) methods estimates DOA by maximizing a spatial spectrum [16].And Fig. 1 illustrates that near end-fire the dominant SRP peak becomes flatter, increasing susceptibility to perturbation-induced peak shifts.Motivated by the section 2, an endfire-targeted strategy was developed:(i) increases the reliability of the time–frequency (T–F) evidence used to form SRP scores, (ii) emphasizes geometrically and spectrally informative cues when they are reliable (longer baselines and higher frequencies), and (iii) stabilizes the nonlinear inversion by fusing in the cosine domain, where the mapping from delay to $\cos\theta$ is linear for ULAs.

Accordingly, we propose a PHAT-based pipeline with two complementary estimators used in our evaluation: **W-SRP-PHAT** performs a reliability-, frequency-, and baseline-weighted SRP search to mitigate peak flattening and reduce peak displacement; **GCC-WLS** refines pairwise delays via oversampled, physically constrained GCC-PHAT and performs a single *cosine-domain* weighted least-squares (WLS) fusion, avoiding repeated ill-conditioned angle inversions near end-fire.

3.1. Sparsity-Aware Frame Selection

In home/office environments, silence frames and late reverberation tails often yield unstable inter-channel phase cues, which directly increases the effective delay uncertainty in (6). We therefore select direct-path-dominant frames using a simple energy gate on a reference channel by accumulating in-band STFT energy (in the spirit of VAD-style frame selection [17]):

$$E(\ell) = \sum_{f\in\mathcal{B}} |Y_1(f,\ell)|^2, \quad \mathcal{B} = [f_{\min}, f_{\max}], \tag{7}$$

and keep the top fraction of frames according to $E(\ell)$. This increases usable SNR and reduces the impact of late reverberation for the subsequent pairwise statistics.

3.2. W-SRP-PHAT: Inverse-Variance Inspired Weighting

To mitigate the variance amplification identified in Section 2, we design a weighting scheme that prioritizes reliable spectral and geometric components. Recall from Eq. (6) that the estimation variance is theoretically proportional to d_p^{-2}. Following the principle of *inverse-variance weighting* (which yields the maximum-likelihood fusion rule for Gaussian errors [15]), the optimal fusion weights should satisfy $w_p \propto d_p^2$. We generalize this by defining the baseline reliability weight as:

$$w_p = \left(\frac{d_p}{d_{\max}}\right)^{\eta}, \quad w_f = \left(\frac{f}{f_{\max}}\right)^{\alpha}, \tag{8}$$

where $\eta \approx 2$ corresponds to the theoretical inverse-variance optimum, while α emphasizes higher frequencies that offer finer phase resolution.

Using a generalized PHAT-β term $\Psi_p(f) = C_p(f)/(|C_p(f)| + \epsilon)^{\beta}(C_p(f)$ indicates the pairwise cross-spectrum),we compute the wideband SRP score:

$$P_{\mathrm{WSRP}}(\theta) = \sum_{f\in\mathcal{B}} \sum_{p\in\mathcal{P}} w_p\, w_f\, \kappa_p(f)^{\gamma} \Re\left\{\Psi_p(f)\, e^{j2\pi f\, \tau_p(\theta)}\right\}, \tag{9}$$

and estimate $\widehat{\theta} = \arg\max_{\theta\in\Theta} P_{\mathrm{WSRP}}(\theta)$. Here, PHAT-$\beta$ can be interpreted as *partial whitening* that interpolates between SRP and SRP-PHAT, and [18-20] has been analyzed for robustness under noise and reverberation. In addition, the coherence term $\kappa_p(f)$ acts as a reliability cue and is related to coherence-weighted SRP-type constructions [16, 21]. This weighted construction directly targets the variance-driven peak displacement mechanism near end-fire by downweighting unreliable T–F evidence and upweighting informative geometry/frequency components.

3.3. GCC-WLS: oversampled, physically constrained GCC-PHAT and cosine-domain fusion

While SRP provides a robust global search, end-fire errors can still be triggered by small delay perturbations amplified by the nonlinear inversion. We therefore refine pairwise delays using an *oversampled* GCC-PHAT [14] computed from the same reliability-weighted PHAT-β cross-spectrum (zero-padding/interpolation in the delay domain, as commonly used for sub-sample time-delay estimation [22]), and constrain the search to the physically feasible delay window:

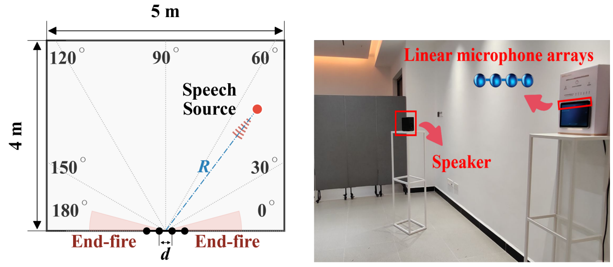
$$|\tau| \le \tau_{\max,p} = \frac{d_p}{c}. \tag{10}$$

Following the standard physical feasibility constraint in TDOA/array processing described in [1], the TDOA estimate is obtained by the dominant peak within this window:

$$\widehat{\tau}_p = \arg \max_{\tau\in[-\tau_{\max,p},\, \tau_{\max,p}]} |r_p(\tau)|\,. \tag{11}$$

Finally, to avoid the ill-conditioned inversion $\theta = \arccos(\cdot)$ near end-fire, we perform fusion in the *linear cosine domain*. For a ULA, the mapping $\tau_p = (d_p/c)u$ (where $u = \cos\theta$) is linear. We extract a pairwise cosine estimate $\hat{u}_p = c\hat{\tau}_p/d_p$ and formulate the fusion as a weighted least squares (WLS) minimization problem:

$$\hat{u} = \arg\min_{u} \sum_{p\in\mathcal{P}} \tilde{w}_p\,(u - \hat{u}_p)^2\,, \tag{12}$$



(a)

(b)

Figure 2: *Experimental environment. (a) Schematic Diagram (far-field:d≪R); (b) On-site Photo.*

where the composite weight $\tilde{w}_p = w_p\bar{\kappa}_p^{\gamma}$ incorporates both geometric leverage (w_p) and signal coherence ($\bar{\kappa}_p$). This is consistent with inverse-variance/WLS fusion principles [15] and with classical LS/WLS formulations in passive (hyperbolic) localization [23, 24]. The closed-form solution to this convex problem is:

$$\hat{u} = \frac{\sum_p \tilde{w}_p\hat{u}_p}{\sum_p \tilde{w}_p}, \quad \hat{\theta} = \arccos(\mathrm{clip}(\hat{u}, -1, 1)). \tag{13}$$

By solving for u before mapping to θ, GCC-WLS maintains statistical stability in the end-fire regions where angular averaging would fail due to the heavy-tailed error distribution.

4. EXPERIMENTS

4.1. Experimental Setup

We conducted experiments in a rectangular room measuring $5\,\mathrm{m} \times 4\,\mathrm{m} \times 3\,\mathrm{m}$, using home-based equipment with an embedded compact linear microphone array. The array comprised four microphones spaced $3.5\,\mathrm{cm}$ apart (total aperture: $10.5\,\mathrm{cm}$), deployed at a height of $1.5\,\mathrm{m}$ and positioned $0.2\,\mathrm{m}$ from the wall.

The sound source playing clean voice signals was placed on the horizontal plane aligned with the linear array. We evaluated the azimuth range of $20°$–$160°$ in $10°$ increments at distances of $1\,\mathrm{m}$ and $2\,\mathrm{m}$. **For each experimental condition (distance and azimuth), we authentically recorded approximately one to two hundred non-overlapping 1-second speech segments.**This environment includes household noise and moderate echo, with a reverberation time similar to that found in an average home and an SNR of 10–20 dB. The specific scene is illustrated by Fig. 2. We define the end-fire region as $\theta_{\mathrm{EF}} = [20°, 40°] \cup [140°, 160°]$.

Real-recorded audio was resampled to 16 kHz. An STFT was performed using a window length of 1024 with 75% overlap (Kaiser window). Frame selection retained the top 30% energy frames in the 800–4500 Hz band, which is closely related to energy/statistical voice activity detection (VAD) ideas [17]. Unless stated otherwise, SRP scanning used a $0°$–$180°$ grid with $0.2°$ resolution. For reliability weighting, we set the frequency exponent to $\alpha = 2$ and the coherence exponent to $\gamma = 1$. For GCC-WLS, GCC-PHAT delay estimation used $16\times$ oversampling with the standard PHAT exponent (i.e., exponent $= 1$). We evaluate two baselines, SRP-PHAT and SRP-MVDR, and two proposed methods, W-SRP-PHAT and GCC-WLS.

4.2. Experimental Results

We report Root Mean Square Error (RMSE) [33], percentage of samples with absolute error less than $6°$ (ACC($6°$)), and the estimation error distribution at 1 m in Fig. 3. Panel (c) shows

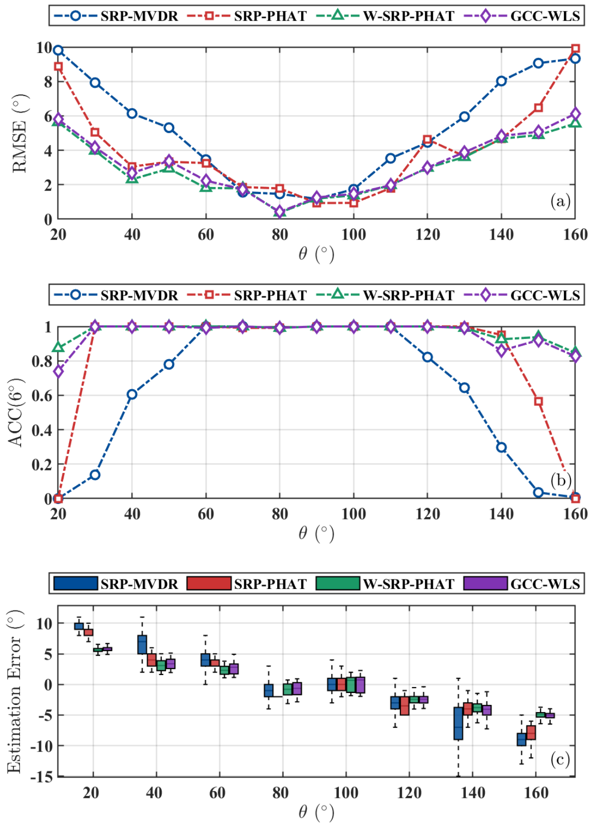


Figure 3: *Localization RMSE (a), ACC (b) and estimation error distribution (c) at a source distance of 1 m under baseline SRP-PHAT, SRP-MVDR and the proposed endfire-robust PHAT methods.The results for each angle are calculated from* ***approximately 120 real-recorded audio data segments****.*

Table 1: *Performance Summary.*

Method	$E_{\mathrm{all}}(^\circ)$	$E_{\mathrm{EF}}(^\circ)$	$\mathrm{ACC}_{\mathrm{EF}}(6^\circ)$	Deg. Span(°)
		at 1m source distance		
SRP-MVDR	5.71	9.29	12.5	50
SRP-PHAT	4.26	6.54	45.5	20
W-SRP-PHAT	**3.21**	**4.71**	**93.0**	**0**
GCC-WLS	**3.49**	**5.20**	**89.1**	**0**
		at 2m source distance		
SRP-MVDR	7.16	10.81	16.2	50
SRP-PHAT	6.20	9.27	20.1	40
W-SRP-PHAT	**5.95**	**6.68**	**58.5**	**0**
GCC-WLS	**5.09**	**6.96**	**45.8**	**0**

a clear shrinkage toward side-fire: errors tend to be positive for $\theta < 90^\circ$ and negative for $\theta > 90^\circ$, indicating a bias toward 90° that becomes more pronounced near end-fire. Relative to the baselines (SRP-PHAT and SRP-MVDR), the proposed W-SRP-PHAT and GCC-WLS exhibit a milder RMSE increase and maintain higher accuracy in the end-fire region.

At 2 m (Fig. 4), estimation bias and RMSE generally increase and accuracy decreases, consistent with more challenging acoustic conditions at longer distance. W-SRP-PHAT and GCC-WLS maintains smaller end-fire errors and a reduced degraded span compared with the baselines.

The overall performance of the proposed methods compared to the baseline methods is summarized in Table 1. We use four key metrics for this quantitative analysis: the average RMSE (E) across all tested azimuths, the end-fire RMSE (E_{EF}), which specifically averages the error within the more challenging end-fire regions ($\theta_{\mathrm{EF}} = [20^\circ, 40^\circ] \cup [140^\circ, 160^\circ]$) and the average accuracy. We introduce the Degraded Span(Deg.Span), which quantifies the total angular width where the ACC(6°) falls below a 30% threshold, indicating regions of unreliable performance.And comparisons between the proposed strategies and baseline methods further demonstrate the enhanced approach's suppression of end-shot effects.

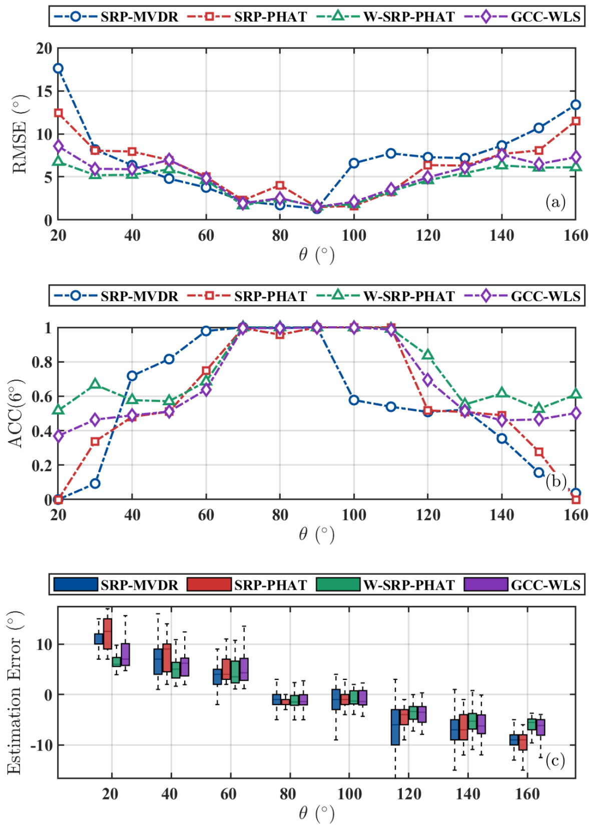


Figure 4: *Experimental results coresponding at a source distance of 2 m.*

5. Conclusion

This paper examined the end-fire degradation problem in far-field DOA estimation using compact uniform linear microphone arrays. We shown that the pronounced error increase near end-fire directions originates from the ill-conditioned nonlinear mapping between TDOA estimates and azimuth angles, which amplifies small delay perturbations into large angular errors. A real-world dataset recorded with a compact linear array in a reverberant indoor environment was used to benchmark several widely used DOA estimation methods, revealing consistent and severe performance degradation near end-fire directions. Based on these analyses, an improved DOA estimation strategy was proposed to mitigate end-fire degradation without increasing array aperture or relaxing the far-field assumption. Experimental results demonstrated improved localization accuracy and stability near end-fire directions while maintaining competitive performance over the full azimuth range.

6. References

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[2] J. Ren, Y. Tian, B. Liu, T. Wu, W. Liu, K.-K. Wong, K.-F. Tong, and K.-M. Luk, “An Fluid Antenna Array-Enabled DOA Estimation Method: End-Fire Effect Suppression,” *IEEE Trans. Veh. Technol.*, vol. 74, no. 11, pp. 13245–13259, Nov. 2025.

[3] E. Grinstein, M. Brookes, and P. A. Naylor, “Graph Neural Networks for Sound Source Localization on Distributed Microphone Networks,” in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[4] R. Takeda and K. Komatani, “Sound source localization based on deep neural networks with directional activate function exploiting phase information,” in *Proc. IEEE ICASSP*, 2016, pp. 405–409.

[5] A. Plinge, F. Jacob, and R. Haeb-Umbach, “Acoustic Microphone Geometry Calibration for Compact Arrays,” *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3151–3163, Jun. 2016.

[6] F. Jacob and R. Haeb-Umbach, “On the Bias of Direction of Arrival Estimation Using Linear Microphone Arrays,” in *ITG-Fachbericht 267*, 2016, pp. 95–99.

[7] Z. Šarić, *et al.*, “Performance Analysis of MVDR Beamformer Applied on an End-fire Microphone Array,” *Arch. Acoust.*, vol. 46, no. 4, pp. 611–621, 2021.

[8] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, pp. 157–180.

[9] H. Do, H. F. Silverman, and Y. Yu, “A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array,” in *Proc. IEEE ICASSP*, vol. 1, 2007, pp. 121–124.

[10] M. Cobos, A. Marti, and J. J. Lopez, “A modified SRP-PHAT functional for robust sound source localization,” *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, Jan. 2011.

[11] J. Capon, “High-Resolution Frequency-Wavenumber Spectrum Analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[12] R. O. Schmidt, “Multiple Emitter Location and Signal Parameter Estimation,” *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[13] H. L. Van Trees, *Optimum Array Processing*. New York, NY, USA: Wiley, 2002.

[14] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[16] E. Grinstein *et al.*, “Steered Response Power for Sound Source Localization: a tutorial review,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, art. no. 59, Nov. 2024, doi: 10.1186/s13636-024-00377-z.

[17] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999, doi: 10.1109/97.736233.

[18] K. D. Donohue, J. Hannemann, and H. G. Dietz, “Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments,” *Signal Processing*, vol. 87, no. 7, pp. 1677–1691, 2007, doi: 10.1016/j.sigpro.2007.01.013.

[19] K. D. Donohue and H. G. Dietz, “Audio signal delay estimation using partial whitening,” in *Proc. IEEE SoutheastCon*, 2007, pp. 466–471, doi: 10.1109/SECON.2007.342946.

[20] H. He, X. Wang, Y. Zhou, and T. Yang, “A steered response power approach with trade-off prewhitening for acoustic source localization,” *J. Acoust. Soc. Am.*, vol. 143, no. 2, pp. 1003–1007, 2018, doi: 10.1121/1.5024652.

[21] P. Nie, B. Liu, P. Chen, and Y. Han, “Coherence-Weighted Steered Response Power for Acoustic Source Localization,” *Acoustics Australia*, vol. 50, pp. 365–371, 2022, doi: 10.1007/s40857-022-00268-3.

[22] J. Benesty, J. Chen, and Y. Huang, “Time-delay estimation via linear interpolation and cross correlation,” *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004, doi: 10.1109/TSA.2004.833008.

[23] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, “Real-time passive source localization: A practical linear-correction least-squares approach,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001, doi: 10.1109/89.966097.

[24] Y. T. Chan and K. C. Ho, “A simple and efficient estimator for hyperbolic location,” *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, Aug. 1994, doi: 10.1109/78.301830.

[25] P. Pertilä, M. Korhonen, and A. Visa, “Phase-based frequency-domain direction of arrival estimation with amplitude weighting,” in *Proc. IEEE ICASSP*, 2010, pp. 2602–2605.

[26] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, “A closed-form method for finding source locations from microphone-array time-delay estimates,” in *Proc. IEEE ICASSP*, vol. 5, 1995, pp. 3019–3022.

[27] Y. Huang, J. Benesty, and G. W. Elko, “Passive acoustic source localization for video camera steering,” in *Proc. IEEE ICASSP*, vol. 2, 2000, pp. 909–912.

[28] Z. Li, Y. Zhang, L. Chen, and H. Zhang, “HearLoc: Locating Unknown Sound Sources in 3D with a Small-Sized Microphone Array,” *IEEE Trans. Mob. Comput.*, vol. 23, no. 12, pp. 16890–16903, Dec. 2024.

[29] T. Long, J. Chen, G. Huang, J. Benesty, and I. Cohen, “Acoustic Source Localization Based on Geometric Projection in Reverberant and Noisy Environments,” *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 1, pp. 143–155, Mar. 2019.

[30] J. Chen, G. Huang, and J. Benesty, “Robust TDOA Estimation for Compact Microphone Arrays in Reverberant Environments,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, no. 5, pp. 1123–1136, May 2020.

[31] D. Salvati, C. Drioli, and G. L. Foresti, “Incoherent Frequency Fusion for Broadband Steered Response Power Algorithms in Noisy Environments,” *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 581–585, May 2014.

[32] X. Wang, G. Huang, J. Benesty, J. Chen, and I. Cohen, “Time Difference of Arrival Estimation Based on a Kronecker Product Decomposition,” *IEEE Signal Process. Lett.*, vol. 28, pp. 51–55, 2021.

[33] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

7. Acknowledgments

Acknowledgments should be included only in the camera-ready version, not in the version submitted for review. For regular papers, pages 5 and 6, and for long papers, pages 9 and 10, are reserved exclusively for acknowledgments, disclosures of the use of generative AI tools, and references. No other content may appear on these pages. Any appendices must be contained within the first four pages for regular papers and within the first eight pages for long papers.

Acknowledgments and references may begin on an earlier page if space permits.

The Interspeech 2026 organizers would like to thank ISCA and the organizing committees of past Interspeech conferences for their help and for kindly providing the previous version of this template.

8. Generative AI Use Disclosure

The extent of Generative AI use must be disclosed. This section may be in the 5th or 6th pages of regular papers, or the 9th or 10th pages of long papers. ISCA policy says: *All (co-)authors must be responsible and accountable for the work and content of the paper, and they must consent to its submission. Any generative AI tools cannot be a co-author of the paper. They can be used for editing and polishing manuscripts, but should not be used for producing a significant part of the manuscript*.