LEAST-SQUARES DOA ESTIMATION WITH AN INFORMED PHASE UNWRAPPING AND FULL BANDWIDTH ROBUSTNESS

Alexander Bohlender[1], *Ann Spriet*[2], *Wouter Tirry*[2], *Nilesh Madhu*[1]

[1] IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
[2] NXP Semiconductors, Product Line Voice and Audio Solutions, Leuven, Belgium

ABSTRACT

The weighted least-squares (WLS) direction-of-arrival estimator that minimizes an error based on interchannel phase differences is both computationally simple and flexible. However, the approach has several limitations, including an inability to cope with spatial aliasing and a sensitivity to phase wrapping. The recently proposed phase wrapping robust (PWR)-WLS estimator addresses the latter of these issues, but requires solving a nonconvex optimization problem. In this contribution, we focus on both of the described shortcomings. First, a conceptually simpler alternative to PWR is presented that performs comparably given a good initial estimate. This newly proposed method relies on an unwrapping of the phase differences vector. Secondly, it is demonstrated that all microphone pairs can be utilized at all frequencies with both estimators. When incorporating information from other frequency bins, this permits a localization above the spatial aliasing frequency of the array. Experimental results show that a considerable performance improvement is possible, particularly for arrays with a large microphone spacing.

Index Terms— sound source localization, direction-of-arrival, least-squares, phase wrapping, spatial aliasing

1. INTRODUCTION

The problem of acoustic source localization is relevant for many practical applications, most notably the enhancement of a desired source by spatial filtering and the separation of several desired sources in the presence of background noise and interferers [1, 2, 3, 4, 5]. An overview of classical methods for direction-of-arrival (DOA) estimation can, for example, be found in [6]. One well-known approach is steered response power with phase transform (SRP-PHAT) [7]. Although originally proposed as a broadband method, it can be used in a narrowband fashion as well, e. g., [8]. Doing so makes it possible to exploit the sparsity of speech over time and frequency [9] for a robust localization of concurrently active sources. However, SRP-PHAT requires searching over a discrete grid. As a result, there is inherently a trade-off between the achievable resolution and the computational complexity.

For resource constrained devices working in realtime, it would be preferable to have an estimator that permits a direct computation of the DOA with low computational effort requirements. One such method is the weighted least-squares (WLS) DOA estimator proposed in [10] that is based on interchannel phase differences. An advantage of this approach over some other closed-form narrowband estimators is that it imposes no restrictions on the array geometry. However, it suffers from several limitations: As it cannot handle ambiguities induced by spatial aliasing, the number of usable microphone pairs decreases with increasing frequency until, eventually, DOA estimation is not possible anymore. Even below this frequency, an increased error is observed due to the problem of phase wrapping. The very recently proposed phase wrapping robust (PWR) estimator [11] can cope with this, but relies on solving a modified problem that is no longer convex. This entails the application of an iterative method and compromises the simplicity of the original approach.

In this contribution, we therefore propose an alternative that does not require changing the least-squares (LS) cost function. Although it is also beneficial to apply this method iteratively, a converged state is reached quickly without the requirement of a manually set convergence threshold. Despite its simplicity, the performance of the newly proposed scheme is similar to PWR at least when sufficiently accurate initial estimates are used. Moreover, both the PWR estimator and the proposed approach can benefit from the availability of good initial estimates as this renders the exclusion of microphone pairs unnecessary. This makes a wider range of applications possible, e. g., when using arrays with a larger microphone spacing.

The remainder of the paper is organized as follows: After a brief description of the employed signal model in Sec. 2, the WLS and PWR-WLS estimators are reviewed in Sec. 3. The deficiencies of the WLS estimator are summarized in Sec. 4 before possibilities to address these, including the proposed informed phase unwrapping (IPU)-LS approach, are discussed. Finally, an experimental evaluation is conducted in Sec. 5, followed by the conclusion in Sec. 6.

2. SIGNAL MODEL

The short-time Fourier transform (STFT) representation of all signals is considered where $\mu = 0, \dots, M-1$ denotes the frequency index and λ the time frame index. For an array of N microphones at positions $\mathbf{r}_1, \dots, \mathbf{r}_N$, the contribution of the desired signal to the microphone signals is written in vector notation as $\mathbf{Y}_s(\mu,\lambda) = [Y_{s,1}(\mu,\lambda), \dots, Y_{s,N}(\mu,\lambda)]^T$. When available in an isolated form, it is possible to infer the DOA and therefore the source position from $\mathbf{Y}_s(\mu,\lambda)$. However, the microphones also capture an additive noise that is, as is done in [10] but also often, e. g., for subspace based approaches like [12, 13], assumed to be spatially white and uncorrelated with the target signal component. It is represented by the vector $\mathbf{V}(\mu,\lambda) = [V_1(\mu,\lambda), \dots, V_N(\mu,\lambda)]^T$, the entries of which are consequently mutually uncorrelated. The result of the additive mixing is the microphone signal vector

$$\mathbf{Y}(\mu,\lambda) = \mathbf{Y}_s(\mu,\lambda) + \mathbf{V}(\mu,\lambda). \tag{1}$$

While more than one source may be active at any given time, it is assumed that only a single source contributes to $\mathbf{Y}_s(\mu,\lambda)$ at one specific (μ,λ). This is in line with the property of W-disjoint orthogonality [9], a common assumption that is typically justified for speech. Furthermore, the widespread model of a single plane wave for the contribution of the target signal is employed. Sufficient accuracy is ensured for the far-field scenario when the direct path dominates over reflections from other directions. With the wavenumber

$$\kappa(\mu) = \frac{2\pi}{c} f_s \frac{\mu}{M} \tag{2}$$

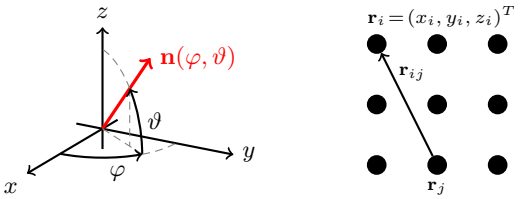
and the DOA vector

$$\mathbf{n}(\varphi,\vartheta) = \begin{bmatrix} \cos(\varphi)\cos(\vartheta) & \sin(\varphi)\cos(\vartheta) & \sin(\vartheta) \end{bmatrix}^T \tag{3}$$

for a coordinate system with azimuth angle $0 \leq \varphi < 2\pi$ and elevation angle $-\frac{\pi}{2} \leq \vartheta < \frac{\pi}{2}$ as depicted in Fig. 1a, this implies that

$$\frac{\mathbf{Y}_s(\mu,\lambda)}{Y_{s,1}(\mu,\lambda)} = \begin{bmatrix} 1 & \mathrm{e}^{\jmath\kappa(\mu)\mathbf{r}_{21}^T\mathbf{n}(\varphi,\vartheta)} & \dots & \mathrm{e}^{\jmath\kappa(\mu)\mathbf{r}_{N1}^T\mathbf{n}(\varphi,\vartheta)} \end{bmatrix}^T. \tag{4}$$

In these expressions, c denotes the speed of sound, f_s the sampling rate and $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ is the position difference between the microphones as illustrated in Fig. 1b.



(a) DOA vector (b) position difference vector

Fig. 1: Illustration of the DOA vector in the coordinate system and the position difference $\mathbf{r}_{ij}$ between two microphones (●)

Because an independent processing of all frames and frequencies is possible, the indices μ and λ will be dropped from the notation and only reintroduced where needed.

3. PRIOR WORK

The original WLS estimator is reviewed in Sec. 3.1. One of its limitations, the phase wrapping problem, as well as the solution proposed for it in [11] are discussed in Sec. 3.2.

3.1. Weighted Least-Squares (WLS) DOA Estimator

In [10], a closed-form solution for the DOA estimation problem is derived that is based on a WLS matching between the observed and expected phases of the power spectral density (PSD) matrix

$$\mathbf{\Phi}_{yy} = E\left\{\mathbf{Y}\mathbf{Y}^H\right\} = \mathbf{\Phi}_{y_s y_s} + \mathbf{\Phi}_{vv}. \tag{5}$$

The assumption of spatially uncorrelated noise implies that the corresponding PSD matrix $\mathbf{\Phi}_{vv}$ is diagonal. Ideally, the off-diagonal elements of $\mathbf{\Phi}_{yy}$ are therefore equal to those of the target signal PSD matrix $\mathbf{\Phi}_{y_s y_s}$. It follows with (4) that the phases of these entries are given by

$$\angle\phi_{yy,ij} = \kappa\, \mathbf{r}_{ij}^T \mathbf{n}(\varphi, \vartheta), \quad 1 \le i, j \le N,\ i \neq j. \tag{6}$$

For an assumed source DOA (φ, ϑ), the corresponding *expected* interchannel phase differences for all unique pairs of microphones are aggregated in a vector $\boldsymbol{\phi}(\mathbf{n})$. In terms of the PSD matrix (5), this vector is composed of the phases of all $(N\,(N-1)\,/2)$ elements above the main diagonal. With (6), it is therefore possible to express the vector of expected phase differences as the product

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{n}) &= \underbrace{\kappa \left[\, \mathbf{r}_{12}\ \mathbf{r}_{13}\ \mathbf{r}_{23}\ \cdots\ \mathbf{r}_{N-1,N} \,\right]^T}_{\mathbf{Q}} \quad \mathbf{n}(\varphi, \vartheta) \\ &= \mathbf{Q} \quad \mathbf{n}(\varphi, \vartheta). \end{aligned} \tag{7}$$

As an estimate of $\boldsymbol{\phi}(\mathbf{n})$, the vector of *observed* phase differences $\widehat{\boldsymbol{\phi}}$ can be extracted from the same elements of the estimated PSD matrix $\widehat{\mathbf{\Phi}}_{yy}$. Given this vector of observed phase differences, the DOA can be estimated by minimizing the least-squares (LS) error with respect to the expected phase differences (7). This yields

$$\widehat{\mathbf{n}} = \mathbf{Q}^{\dagger}\widehat{\boldsymbol{\phi}} \tag{8}$$

where $(\cdot)^{\dagger}$ is the Moore-Penrose pseudoinverse. More generally, when introducing a diagonal weighting matrix $\mathbf{W}$, the DOA vector that minimizes the weighted least-squares (WLS) error is given by

$$\widehat{\mathbf{n}} = (\mathbf{W}\mathbf{Q})^{\dagger}\mathbf{W}\widehat{\boldsymbol{\phi}}. \tag{9}$$

This weighting is used in [10] to exclude microphone pairs that are affected by spatial aliasing, i. e., where the interchannel phase differences (6) are not confined to a range of 2π. From (6) and (2), it follows that this is the case for a pair of microphones with indices i and j at discrete frequencies μ where

$$\mu \ge \frac{M}{f_s}\frac{c}{2\left\|\mathbf{r}_{ij}\right\|_{\ell_2}}. \tag{10}$$

The corresponding entry of the diagonal weighting matrix $\mathbf{W}$ for this frequency is then set to 0, otherwise it is 1.

Regardless of the adopted approach, the estimates of the azimuth and elevation angles can finally be extracted from the DOA vector as

$$\widehat{\varphi} = \operatorname{arctan2}\left(\widehat{n}_y, \widehat{n}_x\right) \tag{11a}$$

$$\widehat{\vartheta} = \arcsin\left(\frac{\widehat{n}_z}{\left\|\widehat{\mathbf{n}}\right\|_{\ell_2}}\right) \tag{11b}$$

where $\operatorname{arctan2}$ is the four-quadrant inverse tangent.

3.2. Phase Wrapping Problem

In [11], it is observed that there is already an increased error directly below the spatial aliasing frequencies. For example, when the expected phase difference is 0.9π, it is possible that due to noise, the phase difference is 1.2π instead. Because phase can only unambiguously be resolved on the range from $-\pi$ to π, this results in an observed phase difference of $1.2\pi - 2\pi = -0.8\pi$. While the error between expected and observed value should actually be 0.3π, an error of 1.7π is seen instead. Likewise, whenever the phase difference is near $\pm\pi$, there is a possibility that the error between expected and observed value is falsely interpreted as being greater than π. As LS generally exhibits a strong sensitivity to outliers since the error is considered in a squared sense, phase wrapping is a critical problem.

To cope with this, [11] proposes to use a modified cost function

$$\widehat{\mathbf{n}} = \arg\min_{\varphi,\vartheta} \left\| \exp\left(\jmath\mathbf{W}\widehat{\boldsymbol{\phi}}\right) - \exp\left(\jmath\mathbf{W}\mathbf{Q}\mathbf{n}(\varphi,\vartheta)\right)\right\|_{\ell_2}^2 \tag{12}$$

that is based on the complex phasors of the phase differences. This solution is referred to as the PWR-WLS estimator in the following. Unlike the original WLS solution, the 2π-periodicity is correctly accounted for. However, with the problem given by (12) being nonconvex, an iterative procedure is required to find $\widehat{\mathbf{n}}$. For the initialization, a reasonably good initial guess such as the result of the original WLS estimator (9) is needed as well.

For benchmarking, [11] introduces an oracle phase unwrapping (OPU) as an upper bound for the performance. Due to the 2π-periodicity of phase, it should not be possible for the elementwise deviation between the vectors of observed and expected phase differences to exceed π. This can be enforced by adding integer multiples of 2π to the entries of the vector according to

$$\widehat{\boldsymbol{\phi}}_{\mathrm{OPU}} = \widehat{\boldsymbol{\phi}} - 2\pi\left\lfloor \frac{\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}(\mathbf{n}_{\mathrm{orc}}) + \pi}{2\pi}\right\rfloor. \tag{13}$$

Elementwise rounding down to the next integer is denoted by $\lfloor\cdot\rfloor$ and $\mathbf{n}_{\mathrm{orc}}$ is the oracle DOA vector, assuming perfect knowledge of the true azimuth and elevation angles.

4. FULL ARRAY AND BANDWIDTH LS ESTIMATOR

We identify the following three problems that limit the usefulness of the original WLS DOA estimator:

(i) The phase wrapping problem [11] causes an increased error near the spatial aliasing frequencies.
(ii) The number of included microphone pairs must be reduced with increasing frequency to avoid spatial aliasing effects.
(iii) An estimation for frequencies at which all microphone pairs are affected by spatial aliasing is not possible at all.

With the PWR-WLS estimator, [11] addresses only issue (i). Nonetheless, we note that it can potentially deal with (ii) as well since the modified problem (12) takes the 2π-periodicity correctly into account. While local optima related to spatial aliasing must be expected, it may still be possible to find the desired solution when a sufficiently accurate initial estimate is available. Likewise, the correction of the phase differences realized by the OPU eliminates not only phase wrapping outliers but also those that are the result of spatial aliasing. This would make the exclusion of microphone pairs used for WLS in [10] as well as for PWR-WLS and OPU-WLS in [11] unnecessary. The resulting schemes for which the weighting matrix $\mathbf{W}$ is set to the identity matrix $\mathbf{I}$ will be referred to as PWR-LS and OPU-LS.

In this work, for addressing all of the aforementioned limitations, we additionally propose an extension of the original WLS estimator [10] that functions as an alternative to the PWR estimator [11]. In Sec. 4.1, it is discussed how this alternative can cope with both (i) and (ii). As shown in Sec. 4.2, both PWR-LS and the newly proposed approach can easily be extended to also addressing (iii), although this requires additional information. As will be demonstrated, data from other frequency bins can be used for this purpose.

4.1. Informed Phase Unwrapping

The OPU (13) offers a simple way to extend the range of phase differences that can be observed beyond 2π. While the oracle DOA vector $\mathbf{n}_{\mathrm{orc}}$ is not available in practice, it is possible to replace it with an initial estimate $\widehat{\mathbf{n}}$. The corrected phase differences are then given by

$$\widehat{\boldsymbol{\phi}}_{\mathrm{IPU}} = \widehat{\boldsymbol{\phi}} - 2\pi \left\lfloor \frac{\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}(\widehat{\mathbf{n}}) + \pi}{2\pi} \right\rfloor . \tag{14}$$

The resulting approach will be termed informed phase unwrapping (IPU) in the following. In this section, it is assumed that the original WLS estimator (9) is used to acquire the initial estimate. Subsequently, a weighting for the exclusion of microphone pairs is no longer needed provided that this initial estimate is sufficiently good. The newly found updated DOA vector estimate can be reinserted into (14) to check whether a different unwrapping is required. Overall, this results in the following iterative procedure:

$$\widehat{\boldsymbol{\phi}}_{\mathrm{IPU}}^{(i)} = \widehat{\boldsymbol{\phi}} - 2\pi \left\lfloor \frac{\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\left(\widehat{\mathbf{n}}_{\mathrm{IPU}}^{(i-1)}\right) + \pi}{2\pi} \right\rfloor \tag{15a}$$

$$\widehat{\mathbf{n}}_{\mathrm{IPU}}^{(i)} = \mathbf{Q}^{\dagger}\widehat{\boldsymbol{\phi}}_{\mathrm{IPU}}^{(i)} \tag{15b}$$

For the initialization, $\widehat{\mathbf{n}}_{\mathrm{IPU}}^{(0)} = \widehat{\mathbf{n}}$ is the initial estimate from (9). Because the correction of the observed phase differences is realized by the addition of *integer* multiples of 2π, "perfect" convergence can be expected after a very low number of iterations, i. e., the corrected vector will be exactly the same as in one of the previous iterations. It is therefore unnecessary to manually set a convergence threshold. At least one iteration is required when not all microphone pairs were used to find the initial estimate, i. e., $\mathbf{W} \neq \mathbf{I}$. Beyond that, an additional iteration is needed only when the unwrapping is not the same as in one of the previous iterations. The termination condition that must be fulfilled for some $j < i$ can thus compactly be written as

$$\widehat{\boldsymbol{\phi}}_{\mathrm{IPU}}^{(i)} = \widehat{\boldsymbol{\phi}}_{\mathrm{IPU}}^{(j)} \quad \text{and} \quad \begin{cases} i > 1 & \text{when } \mathbf{W} \neq \mathbf{I} \\ i > 0 & \text{when } \mathbf{W} = \mathbf{I}. \end{cases} \tag{16}$$

4.2. Extension to the Full Frequency Range

Because an exclusion of microphone pairs is no longer necessary, it is possible to extend the IPU-LS and PWR-LS estimators to higher frequencies. However, the original WLS solution cannot be used to obtain an initial estimate at frequencies where all microphone pairs are affected by spatial aliasing. To compensate for this, the restriction to an entirely independent processing of all (μ, λ) is now dropped. In the single-source case, an initial estimate for use in the otherwise unchanged IPU-LS or PWR-LS estimator can then for example be determined as

$$\widehat{\mathbf{n}}^{(0)}(\mu, \lambda) = \underset{\mu' < \mu}{\mathrm{median}} \left\{ \widehat{\mathbf{n}}^{(\infty)}(\mu', \lambda) \right\} \tag{17}$$

by making use of the final estimates $\widehat{\mathbf{n}}^{(\infty)}(\mu', \lambda)$ of the DOA vector for discrete frequencies $\mu' < \mu$ in the same frame λ. The element-wise applied median operation is favored over a simple averaging due to its robustness to outliers.

5. EVALUATION

The evaluation comprises two parts, the setup for both of which is explained, respectively, in Sec. 5.1.1 and Sec. 5.1.2. Subsequently, results are shown and discussed in Sec. 5.2.

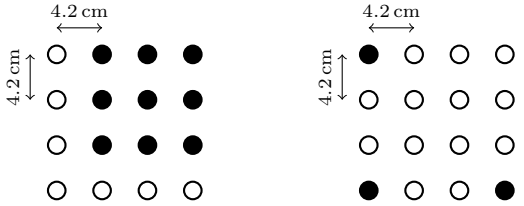
5.1. Setup

The evaluation will be carried out based on impulse responses that were measured using exponential sine sweeps [14] for azimuth angles $\varphi = 0^\circ, 20^\circ, \ldots, 180^\circ$ at a distance of $2\,\mathrm{m}$ between source and array in a meeting room with a reverberation time of about $660\,\mathrm{ms}$.

For the generation of the target signal component, the source signal is convolved with one of the impulse responses at sampling rate $48\,\mathrm{kHz}$ and then downsampled to $16\,\mathrm{kHz}$ where all further processing takes place. The STFT transformation is done on frames of length 512 samples with square-root Hann windows, the frame shift is set to 160 samples. The PSD matrix is estimated through recursive averaging with a time constant of $50\,\mathrm{ms}$ (5 frames). The source and microphone positions are approximately coplanar, so only the estimation of the azimuth angle is considered.

5.1.1. Narrowband error evaluation

First, a synthetic scenario will be used to demonstrate the differences between the different variants of the LS estimator. For this purpose, white noise source signals and temporally uncorrelated noise simulated [15] for a spherically isotropic noise field are used. The noise is mixed additively with the target signal contribution at a signal-to-noise ratio (SNR) of $5\,\mathrm{dB}$. The array is a uniform rectangular array (URA) of $N = 9$ microphones that constitutes a subarray of the miniDSP UMA-16 microphone array [16], as shown in Fig. 2a.



(a) 9-mic subarray (b) 3-mic subarray

Fig. 2: The marked (●) subsets of microphones from the depicted 16-microphone URA [16] are used in the evaluation

The following variants of the LS DOA estimator are compared:

- OPU-LS: (8) with unwrapping given in (13) (upper bound for performance, oracle knowledge required)
- Original WLS [10]: as given in (9)
- PWR-WLS [11]: (12) with initial estimate (9)
  - PWR-LS: the weighting matrix in (12) is set to $\mathbf{W} = \mathbf{I}$
- IPU-LS: (15) with initial estimate (9) (proposed scheme)
  - IPU-LS (early discard): microphone pairs are already excluded at 90% of their aliasing frequency for the initial estimate computation (9) (the factor 0.9 is arbitrary and merely intended as a proof of concept)
- IPU-LS (μ), PWR-WLS (μ) and PWR-LS (μ): proposed variants where the initial estimate is determined as given in (17)

The iterative procedure required for the IPU is terminated upon convergence, which is almost always within 1 or 2 iterations. Although [11] claims that a single iteration already produces good results, we here choose to let PWR proceed until the convergence threshold of 10^{-6} is reached to get an upper bound for its performance.

Two different representations of the angular error will be used for the evaluation. For the first, the absolute error is averaged over $5\,000$ frames and all 10 different angles. Secondly, for facilitating the interpretation, an alternative representation of the error

$$\varepsilon_{\mathrm{rel}} = \frac{\varepsilon - \varepsilon_{\mathrm{OPU}}}{\varepsilon_{\mathrm{WLS}} - \varepsilon_{\mathrm{OPU}}} \tag{18}$$

is considered where ε is the aforementioned mean absolute error, $\varepsilon_{\mathrm{OPU}}$ and $\varepsilon_{\mathrm{WLS}}$ are, respectively, the corresponding errors of the OPU-LS and WLS estimators. This metric therefore describes how an approach compares to the oracle method (lower error bound) and WLS (upper error bound).

5.1.2. Broadband error evaluation

In the second part, the accuracy of the broadband estimates is assessed under more practical conditions. A total of 300 utterances were selected at random from the TSP speech database [17]. For a good diffuseness, the additive noise was recorded by simultaneously playing back slightly delayed versions of the ETSI background noise database [18] pub noise signal on loudspeakers facing each of the 4 corners of the room. Even so, a minor impact of the imperfect diffuseness on the results cannot fully be ruled out. Still, a too strong effect thereof is averted as only those frames are included in the DOA estimation where the SNR is no more than 10 dB below the global mixing SNR.

To demonstrate the usefulness of taking the full frequency range into account, the triangular subarray shown in Fig. 2b is now used as well. Because of the comparatively large microphone spacing, an estimation with the original WLS estimator is only possible for frequencies up to 1361 Hz.

For one frame, the corresponding broadband estimate is chosen as the maximum of the histogram of all valid estimates for frequencies up to 6 kHz. The histogram bins are centered around $\varphi = 0^\circ, 5^\circ, \ldots, 180^\circ$. The percentage of estimates where the absolute error does not exceed a threshold of 0°, 5° or 15° will serve as a measure that is unaffected by the exact position of outliers. As a reference, the results for narrowband SRP-PHAT are also displayed.

5.2. Results

Figure 3 shows the results for the first experiment. The upper part of the figure depicts the averaged error in degrees, the relative error (18) for the frequency range where the phase wrapping effect is most relevant is shown in the lower part. First, it is compared how the methods perform for an entirely independent estimation at all (μ, λ). Clearly, the use of IPU-LS (-+-) leads to an improvement over the original WLS estimator (—). The comparison of PWR-WLS (-●-) with IPU-LS and PWR-LS (-■-) indicates that the omission of the weighting causes a greater sensitivity to initialization errors: PWR-WLS shows the best performance of the three under these conditions, but the PWR-LS error ultimately eclipses even the original WLS error at higher frequencies (above 3.1 kHz). This can be explained as the result of the intensifying spatial aliasing ambiguities. On the other hand, the IPU-LS performance is more consistent overall and, at lower frequencies, is only slightly worse than that of PWR-LS.

One approach to reduce the initial estimate error could be to already exclude microphone pairs before their spatial aliasing frequency is reached (here done for IPU-LS, -◇-). In the frequency range where this early rejection makes a difference, e. g., around 2 kHz and 3 kHz, an improvement is indeed visible.

Next, we consider the variants where lower frequency estimates are used to find the initial estimate. Both IPU-LS (-+-) and PWR-LS (-△-) then come close to OPU-LS (- -) from around 2 kHz on in this example. These are also the only methods that permit a DOA estimation even when all microphone pairs are affected by spatial aliasing. As the comparison with PWR-WLS (-×-) shows, the introduction of a weighting is now no longer beneficial. Moreover, the results for IPU-LS are a little better yet than those for PWR-LS in this case.

The results for the second experiment can be found in Fig. 4. It is indicated at the top of each part of the figure which array and SNR were, respectively, used. For the 3-microphone subarray, there is a substantial improvement when making use of the full frequency range with either the IPU (▮) or the PWR (▮) estimator. It now seems that the results obtained by solving the nonconvex PWR problem are a little better than for the simple IPU-LS. This could be related to the use of recorded diffuse noise as opposed to the simulated noise in the previous experiment. Generally, the differences are more subtle for the 9-microphone subarray because of the greater frequency range where a reasonably large number of microphone pairs is available. Still, the comparison between the approaches qualitatively shows the same relations as for the other subarray. The comparison with SRP-PHAT (▮) indicates that it is possible to achieve an equivalent performance with the LS estimator. On the other hand, the low complexity is an advantage over the former since no grid search is needed.

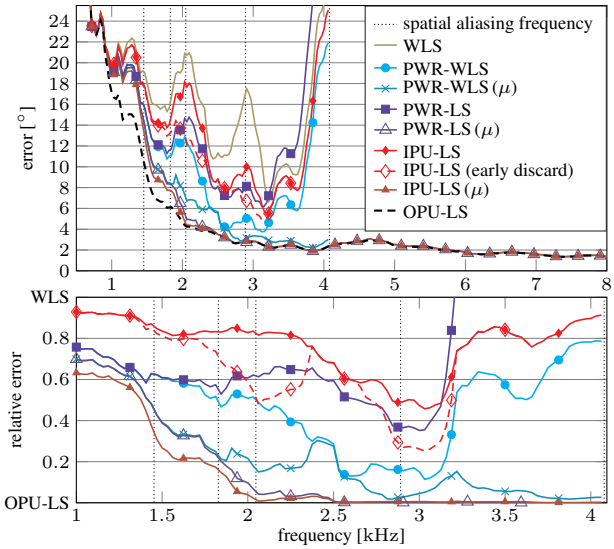


Fig. 3: Results for the setup from Sec. 5.1.1 are best compared in terms of the relative error (18) (bottom plot), note that a moving average of length 10 (312.5 Hz) was applied to smoothen the lines (top and bottom plots) for better illustration of the trends

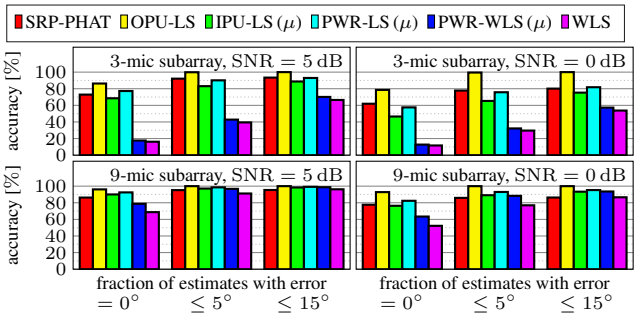


Fig. 4: Results for the setup from Sec. 5.1.2 in terms of the broadband angular error with a resolution of 5°

6. CONCLUSIONS

In this paper, the limitations of the WLS DOA estimation approach are alleviated in two ways. First, a simple alternative to the PWR estimator is proposed for addressing the phase wrapping problem. This method termed IPU typically achieves perfect convergence within 1 or 2 iterations, without requiring a manually set threshold. Secondly, it is demonstrated that, given an initial estimate, it is possible with both the IPU and PWR approach to take all microphone pairs into account regardless of spatial aliasing ambiguities. This allows an estimation even at frequencies where all microphone pairs are excluded in the original WLS estimator. Evaluations are conducted first for synthetic conditions, then on speech with recorded diffuse noise. For both approaches, an improvement can be expected from taking the full bandwidth into account. The comparison with PWR shows that it is possible to get a comparable performance with the newly proposed IPU method despite its simplicity.

The evaluation was restricted to the single-source case here. As long as all processing is done independently for each frame and frequency, the extension of the IPU to the multisource case is straightforward. However, when using DOA estimates from lower frequencies to initialize IPU or PWR, a different initial estimate is required for each of the sources. It is left for future work to explore what the best strategy is under these conditions.

7. REFERENCES

[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, Wiley, 2018.

[2] P. Vary and R. Martin, *Digital Speech Transmission - Enhancement, Coding & Error Concealment*, John Wiley & Sons, Ltd., Jan. 2006.

[3] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2182–2196, Dec 2014.

[4] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1900–1912, Sep. 2011.

[5] S. U. N. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind speech separation and enhancement with GCC-NMF," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 745–755, April 2017.

[6] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds., pp. 135–170. John Wiley & Sons, Ltd., New York, USA, 2008.

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

[8] M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a Laplacian mixture model," *Digital Signal Processing*, vol. 21, no. 1, pp. 66 – 76, 2011.

[9] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2002, vol. 1, pp. I–529–I–532.

[10] O. Thiergart, W. Huang, and E. A. P. Habets, "A low complexity weighted least squares narrowband DOA estimator for arbitrary array geometries," in *Proc. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 340–344.

[11] T. Kabzinski and E. A. P. Habets, "A least squares narrowband DOA estimator with robustness against phase wrapping," in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, Sep. 2019.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, March 1986.

[13] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[14] ISO 18233:2006, "Acoustics — application of new measurement methods in building and room acoustics," Standard, International Organization for Standardization, Geneva, Switzerland, June 2006.

[15] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[16] miniDSP, "UMA-16 USB microphone array," `https://www.minidsp.com/products/usb-audio-interface/uma-16-microphone-array`, Accessed: Oct. 11, 2019.

[17] P. Kabal, "TSP speech database," Tech. Rep., McGill University, Montreal, Quebec, Canada, 2002.

[18] "Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database," *ETSI EG 202 396-1*, 2008.