DOA ESTIMATION OF AUDIO SOURCES IN REVERBERANT ENVIRONMENTS

Jesper Rindom Jensen[1], *Jesper Kjær Nielsen*[2,3], *Richard Heusdens*[1,4], *and Mads Græsbøll Christensen*[1]

[1]Aalborg University, Denmark
Audio Analysis Lab, AD:MT
{jrj,mgc}@create.aau.dk

[2]Aalborg University, Denmark
Dept. of Electronic Systems
jkn@create.aau.dk

[3]Bang & Olufsen A/S
Struer, Denmark

[4]Delft University of Technology, NL
Dept. of Microelectronics
r.heusdens@tudelft.nl

ABSTRACT

Reverberation is well-known to have a detrimental impact on many localization methods for audio sources. We address this problem by imposing a model for the early reflections as well as a model for the audio source itself. Using these models, we propose two iterative localization methods that estimate the direction-of-arrival (DOA) of both the direct path of the audio source and the early reflections. In these methods, the contribution of the early reflections is essentially subtracted from the signal observations before localization of the direct path component, which may reduce the estimation bias. Our simulation results show that we can estimate the DOA of the desired signal more accurately with this procedure compared to state-of-the-art estimator in both synthetic and real data experiments with reverberation.

Index Terms— Audio localization, DOA estimation, reverberation, nonlinear least squares, maximum likelihood

1. INTRODUCTION

One of the key topics within microphone array signal processing is the estimation of the direction-of-arrival (DOA) of an acoustic source in relation to an array of microphones, sometimes briefly referred to as the localization problem. This is not without reason since numerous applications rely on such information. Extracting information about the whereabouts of an acoustic source can be useful in itself for surveillance systems, but the source location is also extremely important if the problem is to separate the source from undesired noise and interference which is a typical scenario in, e.g., hands-free telephony and hearing aids. Due to its importance, the localization problem is therefore a well-established research problem and an enormous body of scientific contributions have dealt with it.

Many existing localization methods developed for, e.g., radar and wireless communication applications assume that the source to be localized is narrowband [1–5], which simplifies the localization problem significantly. This is, however, a poor assumption when dealing with localization of audio sources using microphone arrays, as these are well known to carry information over a wide range of frequencies. Essentially, the localization problem has to be tackled in a different way because of this, and this has resulted in a number of different approaches for localization of such broadband sources. A popular approach, and a direct extension of the narrowband localization problem, is to transform the microphone recordings to the frequency domain. This enables processing of the recordings in subbands, where the aforementioned narrowband methods can be applied [6–8]. After estimation of the location in each subband, these estimates are combined to achieve a location estimate of the broadband source. Another popular procedure is to conduct the localization in two stages. First, the time differences-of-arrival (TDOAs) of the desired source between all possible microphone pairs in the microphone array are estimated [9, 10] after which they are mapped to a location estimate [11, 12]. This approach is computationally efficient, but is mainly targeted towards single-source scenarios. A more recent approach is to exploit a model, e.g., the harmonic model, for the desired source when estimating the location [13–15], which can lead to a higher estimation accuracy in anechoic and mildly reverberant environments [14, 15]. Common for many of these methods is that they do not explicitly take reverberation into account. Reverberation, however, is present in nearly all situations where the localization of audio source is to be estimated. Moreover, it is well known to have a detrimental impact on the performance of localization methods, and thus puts strong limitations on the better part of these. A few of these methods have been claimed to be relatively robust against reverberation, such as the SRP-PHAT method [9, 16], despite the fact that they have not been explicitly derived to tackle this problem.

In this paper, we seek to take a step closer to the solution to localization of broadband, audio sources in reverberant environments. We do this by modeling the reverberation, taking the short-term periodic nature of audio sources into account, and deriving localization methods based on these observations. More specifically, we exploit a model for the so-called early reflections to subtract the contribution from these when estimating the location of the desired source to decrease the estimation bias. This is achieved by adopting a similar estimation procedure to the RELAX method originally proposed in [17]. The proposed approach, although not considered herein, may also be useful for room geometry estimation where the locations of the reflections are needed [18].

In the remaining sections of the paper, we introduce the abovementioned models and the estimation problem (Section 2), derive DOA estimators based on these models (Section 3), provide experimental results (Section 4), and discuss our findings (Section 5).

2. MODEL AND PROBLEM DESCRIPTION

We consider the scenario where the sound field of an acoustic source is sampled by a microphone array in a room. In this particular scenario, the discrete time recording at time instance n obtained by microphone k of the microphone array can be modeled as

$$y_k(n) = (s' * g_k)(n) + v'_k(n) \tag{1}$$

$$= s_k(n) + v'_k(n) \tag{2}$$

where $*$ is the discrete time convolution operator, $s'(n)$ is the clean source signal of interest, $s_k(n) = (s' * g_k)(n)$ is the desired signal

The work by J. R. Jensen was supported by the Danish Council for Independent Research, grant ID: DFF 1337-00084. The work by J. K. Nielsen was supported by InnovationsFonden. The work by R. Heusdens and M. G. Christensen was partly supported by the Villum Foundation.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

including reverberation, $g_k(n)$ is the room impulse response from the desired source to microphone k, $v'_k(n)$ is additive noise, which could be interfering sources or sensor noise, and all the signals are real valued. In this paper, we focus on the problem of robust DOA estimation in reverberant environments, while assuming the noise, $v'_k(n)$, to be white Gaussian. If we have K microphones and record N samples for each microphone, we can model all observations as

$$\mathbf{y} = \begin{bmatrix}\mathbf{y}_1^T & \mathbf{y}_2^T & \cdots & \mathbf{y}_K^T\end{bmatrix}^T = \mathbf{s} + \mathbf{v}', \tag{3}$$

with $\mathbf{y}_k = [y_k(0) \ \cdots \ y_k(N-1)]^T$, and $\mathbf{s}$ and $\mathbf{v}'$ are defined similar to $\mathbf{y}$ but contain instead the desired signal including reverberation and the additive noise respectively. To facilitate the estimation of the DOA of the clean desired signal with respect to the array of microphones, we further specify the model in two ways: 1) we introduce a model for the clean desired signal, and 2) we assume a certain microphone array structure.

Regarding the signal model, we assume that the clean desired signal is periodic which implies that the harmonic model can be assumed, i.e.,

$$s'(n) = \sum_{l=-L, l\neq 0}^{L} \alpha_l e^{jl\omega_0 n}, \tag{4}$$

where $\alpha_{-l} = \alpha_l^*$ is the complex amplitude of the l'th harmonic, ω_0 is the fundamental frequency, and L is the number of harmonics. It is important to note that this has the widely used broadband model for DOA estimation as a special case if we chose $\omega_0 = 2\pi/N$ and $L = N$ [19], and the used model is therefore less restrictive than it might seem at first glance. Regarding the array structure, we assume a uniform linear array (ULA) as a proof-of-concept, but the methods developed in the following could just as well be derived or generalized to other array structures. With the ULA structure and an assumption of the desired source being in the far-field, we know that the r'th source impinging on the array is received at microphone k with a delay of

$$\tau_{r,k} = k\frac{d \sin\theta_r}{c} = k\eta_r, \tag{5}$$

relative to microphone 1. Here, d is the spacing between to adjacent microphones of the array, θ_r is the DOA of the r'th source with respect to the ULA, and c is the sound propagation speed. We then further assume that the desired signal including reverberation can be modeled as a sum of harmonic signals constituting the early reflections, and a noise component $\mathbf{v}$, constituting both the late reverberation and the sensor noise. In this case, we can rewrite the model in (3) as

$$\mathbf{y} = \sum_{r=1}^{R} \mathbf{H}(\eta_r)\boldsymbol{\alpha}_r + \mathbf{v}, \tag{6}$$

with $\mathbf{H}(\eta_r) = [\mathbf{Z}^T \ (\mathbf{Z}\mathbf{D}_2(\eta_r))^T \ \cdots \ (\mathbf{Z}\mathbf{D}_K(\eta_r))^T]^T$, and

$$\mathbf{Z} = \begin{bmatrix}\mathbf{z}_1 & \cdots & \mathbf{z}_L & \mathbf{z}_1^* & \cdots & \mathbf{z}_L^*\end{bmatrix}, \tag{7}$$

$$\mathbf{z}_l = \begin{bmatrix}1 & e^{jl\omega_0} & \cdots & e^{j(N-1)l\omega_0}\end{bmatrix}^T, \tag{8}$$

$$\mathbf{D}_k(\eta_r) = \text{diag}\left(\begin{bmatrix}\mathbf{d}_k^T(\eta_r) & \mathbf{d}_k^H(\eta_r)\end{bmatrix}\right), \tag{9}$$

$$\mathbf{d}_k(\eta_r) = \begin{bmatrix}e^{-j\omega_0 k\eta_r} & \cdots & e^{-jL\omega_0 k\eta_r}\end{bmatrix}^T, \tag{10}$$

and diag$(\cdot)$ is the operator creating a diagonal matrix from a vector. The problem at hand is then to estimate the DOA of the harmonic signal corresponding to $r = 1$, which corresponds to the direct path component.

3. DOA ESTIMATION WITH REVERBERATION

To accurately estimate the DOA of the clean desired signal, we consider the estimation of the DOAs of all the R reflections in (6) as opposed to many state-of-the-art DOA estimators that do not take reverberation into account. The goal of this procedure is to reduce the bias incurred by the early reflections when estimating the DOA of the direct path component. In this paper, we solve this estimation problem by using a nonlinear least squares (NLS) estimator, which equals the maximum likelihood estimator when the noise $\mathbf{v}$ is white Gaussian. That is, our estimator of the DOAs of the R reflections can be written as

$$\{\widehat{\boldsymbol{\eta}}, \widehat{\overline{\boldsymbol{\alpha}}}\} = \arg\min_{\{\boldsymbol{\eta}, \overline{\boldsymbol{\alpha}}\}} \|\mathbf{y} - \overline{\mathbf{H}}(\boldsymbol{\eta})\overline{\boldsymbol{\alpha}}\|_2^2, \tag{11}$$

where $(\widehat{\cdot})$ denotes an estimate of a parameter, $\boldsymbol{\eta} = [\eta_1 \ \cdots \ \eta_R]^T$, and

$$\overline{\mathbf{H}}(\boldsymbol{\eta}) = \begin{bmatrix}\mathbf{H}(\eta_1) & \cdots & \mathbf{H}(\eta_R)\end{bmatrix}, \tag{12}$$

$$\overline{\boldsymbol{\alpha}} = \begin{bmatrix}\boldsymbol{\alpha}_1^T & \cdots & \boldsymbol{\alpha}_R^T\end{bmatrix}^T. \tag{13}$$

We simplify this estimator by solving for the amplitudes, $\overline{\boldsymbol{\alpha}}$, which leads to

$$\widehat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \|\mathbf{P}_{\overline{\mathbf{H}}}^{\perp}\mathbf{y}\|_2^2, \tag{14}$$

where

$$\mathbf{P}_{\mathbf{A}}^{\perp} = \mathbf{I} - \mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H, \tag{15}$$

and $\mathbf{I}$ is the identity matrix. While this estimator should have optimal performance in terms of estimation variance when the noise is white Gaussian, it is extremely time consuming to implement in practice as it is both nonconvex and multidimensional.

To alleviate the computational complexity issue, we propose to implement the estimator similar to the RELAX procedure proposed in [17] and since used in, e.g., [20, 21]. To facilitate the estimation procedure, we introduce the modified observed signal vector

$$\mathbf{y}_r = \mathbf{y} - \sum_{q=1, s\neq r}^{R} \mathbf{H}(\widehat{\eta}_q)\widehat{\boldsymbol{\alpha}}_q, \tag{16}$$

where the estimates $\{\widehat{\eta}_q, \widehat{\boldsymbol{\alpha}}_q\}_{q=1, q\neq r}^{R}$ are assumed to be given. This suggests that we can estimate the η_r and $\boldsymbol{\alpha}_r$ using simpler estimators, i.e.,

$$\widehat{\boldsymbol{\alpha}}_r = (\mathbf{H}^H(\eta_r)\mathbf{H}(\eta_r))^{-1}\mathbf{H}^H\mathbf{y}_r, \tag{17}$$

$$\widehat{\eta}_r = \arg\min_{\eta_r} \|\mathbf{P}_{\mathbf{H}(\eta_r)}^{\perp}\mathbf{y}_r\|_2^2. \tag{18}$$

That is, with these estimators we can obtain disjoint estimates of the DOAs of the R reflections. This leads to the RELAX version of the estimator in (14), which is implemented in the following steps:

Step (1): Assume $R = 1$. Estimate η_1 and $\boldsymbol{\alpha}_1$ from $\mathbf{y}_1 = \mathbf{y}$ as described above.

Step (2): Assume $R = 2$. Estimate η_2 and $\boldsymbol{\alpha}_2$ from $\mathbf{y}_2$ computed using (16) and the parameter estimates obtained in Step (1). Then, re-estimate η_1 and $\boldsymbol{\alpha}_1$ from $\mathbf{y}_1$ computed using the newly obtained estimates of η_2 and $\boldsymbol{\alpha}_2$. Proceed by iterating between these two substeps until "practical convergence".

Step (3): Assume $R = 3$. Estimate η_3 and $\boldsymbol{\alpha}_3$ from $\mathbf{y}_3$ computed using $\{\widehat{\eta}_q, \widehat{\boldsymbol{\alpha}}\}_{q=1}^{2}$ from Step (2). Then, re-estimate η_1 and $\boldsymbol{\alpha}_1$ from $\mathbf{y}_1$ computed using $\{\widehat{\eta}_q, \widehat{\boldsymbol{\alpha}}\}_{q=2,3}$. Also, re-estimate η_2 and $\boldsymbol{\alpha}_2$ from $\mathbf{y}_2$ computed using $\{\widehat{\eta}_q, \widehat{\boldsymbol{\alpha}}\}_{q=1,3}$. Proceed by iterating between these three substeps until "practical convergence".

Remaining steps: Continue similarly to the previous steps until R is equal to the number of early reflections.

To check for convergence, we can compute the cost function in iteration i of each step as

$$J(i) = \|\mathbf{y} - \overline{\mathbf{H}}(\widehat{\boldsymbol{\eta}})\widehat{\boldsymbol{\alpha}}\|_2^2. \tag{19}$$

If the change in the cost function between two consecutive iterations is smaller than a threshold value, i.e., $|J(i) - J(i-1)| < \epsilon$, we proceed to the next step, or terminate the estimation procedure if all reflections have been estimated. We refer to this algorithm as the relaxed NLS estimator (RNLS).

We also propose an alternative estimation procedure for the estimation of the DOAs of the early reflections. If the clean desired signal is stationary, which is a reasonable assumption for short time frames of audio sources, the early reflections are just an attenuated and delayed version of the direct path component. This is not explicitly taken into account in the estimation problem in (11) and, eventually, this may influence on the convergence speed of the RELAX based estimation procedure. We therefore set up the following model based on the aforementioned observations:

$$\mathbf{y} = \sum_{r=1}^{R} \gamma_r \mathbf{H}(\eta_r)\mathbf{T}_r\boldsymbol{\alpha} + \mathbf{v}, \tag{20}$$

where γ_r is the attenuation of reflection r in relation to the direct path component (i.e., $\gamma_1 = 1$),

$$\mathbf{T}_r = \operatorname{diag}\left(\begin{bmatrix}\mathbf{t}_r^T & \mathbf{t}_r^H\end{bmatrix}\right), \tag{21}$$

$$\mathbf{t}_r = \begin{bmatrix}e^{j\omega_0\xi_r} & \cdots & e^{jL\omega_0\xi_r}\end{bmatrix}^T, \tag{22}$$

$\boldsymbol{\alpha}$ is the harmonic amplitudes of the direct path component, and ξ_r is the delay of reflection r in relation to the direct path component (i.e., $\xi_1 = 0$). The NLS estimator of the DOAs of the reflections based on this signal model is given by

$$\{\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\alpha}}\} = \arg \min_{\{\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\alpha}\}} \|\mathbf{y} - \sum_{r=1}^{R} \gamma_r \mathbf{H}(\eta_r)\mathbf{T}_r\boldsymbol{\alpha}\|_2^2. \tag{23}$$

This estimation problem is obviously more complex than the estimator in (11), since 1) it has a higher dimensionality, and 2) we can not easily simplify the estimator because the closed-form estimates of γ_r, for $r = 1, \ldots, R$, and $\boldsymbol{\alpha}$ will depend on each other. To simplify the estimator, we therefore again adopt the RELAX procedure. For this implementation, we have the following modified signal model

$$\mathbf{y}_r = \mathbf{y} - \sum_{q=1, q\neq r}^{R} \widehat{\gamma}_q \mathbf{H}(\widehat{\eta}_q)\widehat{\boldsymbol{\alpha}}. \tag{24}$$

Exploiting this formulation, we obtain a closed-form estimate of $\boldsymbol{\alpha}$ when $r = 1$ as

$$\widehat{\boldsymbol{\alpha}} = [\mathbf{H}^H(\eta_1)\mathbf{H}(\eta_1)]^{-1}\mathbf{H}^H(\eta_1)\mathbf{y}_1, \tag{25}$$

while, for $r = 2, \ldots, R$, the optimal estimate of γ_r is given by

$$\widehat{\gamma}_r = \frac{\operatorname{Re}\{\widehat{\boldsymbol{\alpha}}^H \mathbf{T}_r^H \mathbf{H}^H(\eta_r)\mathbf{y}_r\}}{\widehat{\boldsymbol{\alpha}}^H \mathbf{T}_r^H \mathbf{H}^H(\eta_r)\mathbf{H}(\eta_r)\mathbf{T}_r\widehat{\boldsymbol{\alpha}}}. \tag{26}$$

Moreover, the DOA of the direct path component are estimated as

$$\widehat{\eta}_1 = \arg \min_{\eta_1} \|\mathbf{P}_{\mathbf{H}(\eta_1)}^{\perp}\mathbf{y}_1\|_2^2, \tag{27}$$

while the DOAs and delays of the early reflections are estimated jointly as

$$\{\widehat{\eta}_r, \widehat{\xi}_r\} = \arg \min_{\eta_r, \xi_r} \|\mathbf{y}_r - \widehat{\gamma}_r \mathbf{H}(\eta_r)\mathbf{T}_r\widehat{\boldsymbol{\alpha}}\|_2^2 \tag{28}$$

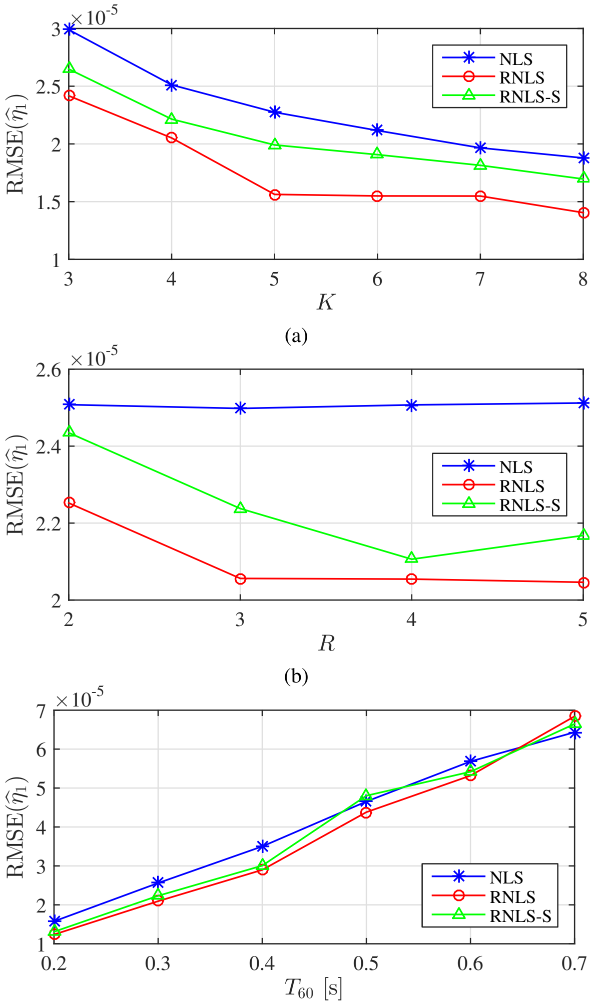
for $r = 2, \ldots, R$. This alternative estimation procedure is then implemented by using the above estimators in an iterative algorithm following the same steps as the RNLS method. We refer to this algorithm as the RNLS estimator with structured amplitudes (RNLS-S). Note that $\widehat{\gamma}_r$ might be negative although this has no physical meaning. To avoid this, a non-negative least squares method can be used for estimation of γ_r [22], but this is not considered herein. In conclusion, we note that both proposed methods are identical for $R = 1$, in which case they therefore resemble the NLS estimator proposed in [14].

Table 1. RMSEs of η_1 estimated from real speech using the NLS, RNLS, RNLS-S, and SRP-PHAT methods.

NLS	RNLS	RNLS-S	SRP-PHAT
$3.8 \cdot 10^{-5}$	$3.6 \cdot 10^{-5}$	$3.6 \cdot 10^{-5}$	$5.4 \cdot 10^{-5}$

4. EXPERIMENTAL RESULTS

We evaluated the proposed methods on both synthetic and real data. In the evaluation on synthetic data, we investigated the performance of the proposed estimators versus the number of microphones, number of assumed reflections, and the reverberation time. The synthetic data, sampled at 8 kHz, was assumed to be a harmonic signal with six, unit amplitude, harmonics, each having a uniformly distributed random phase and the fundamental frequency was 255.2 Hz. The fundamental frequency as well as the number of harmonics were assumed to be known to simplify this evaluation, however, we also include the estimation of these in the evaluation on real data. The signal was then synthesized spatially using a RIR generator [23]. The spatially synthesized signal was generated by using a ULA structure parallel to the x-axis of a room with dimensions $8 \times 6 \times 4$ m and it was centered around [3.5, 1, 1] m. Furthermore, the microphone spacing in the ULA was 5 cm. The other options in the generator was set as follows: the microphone types were omnidirectional, the sound speed was 343 m/s, and high-pass filtering was used. Moreover, white Gaussian noise was added to each channel at an SNR of 40 dB to represent sensor noise. With this setup, we then applied the proposed methods on blocks of $N = 200$ samples from each microphone. In both proposed methods, a tolerance of $\epsilon = 10^{-3}$ was used. However, the maximum number of iterations for each reflection was set to 20 to speed up the processing time. A uniform search grid for the estimation of the η_r's of size 200 was used in the interval $[-\eta_{\max}; \eta_{\max}]$ where $\eta_{\max} = d/c$. Moreover, the search grid for the delays ξ_r in the RNLS-S method was uniform, of size 100, and defined over the interval $[0; 2\pi/\omega_0]$. The simulation



(a)

(b)

(c)

Fig. 1. Plot of the RMSE of the η_1 estimates obtained using the NLS, RNLS, and RNLS-S methods in different scenarios.

was repeated for different angles of the desired source, i.e., for the angles $-80^\circ, -75^\circ, \ldots, 80^\circ$, while the source-to-array-center distance was 2.5 m and the source had the same z-coordinate as the array. The root mean squared error (RMSE) of the estimates of η_1 were then measured over these different angles. In the first evaluation, the T_{60} and the assumed number of early reflections were 0.3 s and 3, respectively, while the number of microphones was varied (Fig. 1a). The proposed methods outperforms the NLS estimator [14], which do not take reverberation into account, for all different numbers of microphones. Interestingly, the RNLS method yields the best performance in all cases. Second, the number of microphones was fixed to four, while the number of assumed early reflections was varied (Fig. 1b). Again, the proposed methods outperforms the NLS method, and, generally, higher R's seems to provide better performance. Finally, the numbers of microphone and assumed reflections were set to four and three, respectively, for different reverberation times (Fig. 1c). For reverberation times smaller than or equal to 0.6 s, the RNLS method outperforms the NLS methods, whereas the RNLS-S needs reverberation times below 0.5 s to outperform NLS.

The proposed methods were also evaluated on real, single-channel data that were spatially synthesized. This was a four seconds long, female speech signal. The signal was single-channel and, therefore, synthesized spatially using the aforementioned RIR generator. In this evaluation, the ULA was centered around [3.5, 1.5, 1] m, while the source was moving from [1, 4, 1.7] m to [7 4 1.7] m over the duration of the sentence. The pitch and the number of harmonics were estimated from the first microphone using the NLS estimator in [24]. Besides this, the number of assumed reflections was 4, the number of microphones was 4, and the reverberation time was 0.3 s, while the remaining setup parameters were identical to the previous evaluations. The proposed estimators as well as the popular SRP-PHAT method were applied on the generated data (Fig. 2). The SRP-PHAT method was implemented by applying an FFT of length 256, and by integrating over frequencies in the interval [100;4000] Hz. The NLS based methods, including the proposed, all seem to provide better η_1 estimates than the SRP-PHAT method in general. We also measured the RMSEs of the estimated η_1's over time (Tab. 1). These support our findings from the synthetic data evaluation that the proposed methods yield better performance than the NLS method.

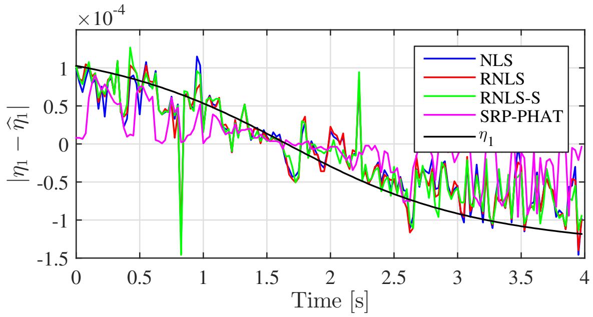


Fig. 2. Absolute errors of the NLS, RNLS, RNLS-S, and SRP-PHAT methods when applied on a real, female speech signal.

5. DISCUSSION

In this paper, we considered the topic of DOA estimation of audio and speech sources in the presence of reverberation. Most existing DOA estimators do not explicitly take into account the reverberation [6–8, 11–14], and are thus negatively influenced by it. A few estimators have been reported to be relatively robust against reverberation though [9, 16]. In this paper, we imposed a model for the early reflections as well as for the audio source. Based on these models, we then proposed two iterative DOA estimation methods that estimate the DOAs of both the direct path component and the early reflections. By doing this, the early reflections are subtracted from the signal observations before estimating the DOA of the direct path component, which should decrease the estimation bias. Our experimental results confirm that the proposed methods can indeed outperform state-of-the-art methods for DOA estimation of audio sources in scenarios with reverberation. While not considered herein, the proposed methods may also be used in room geometry estimation, where the DOAs of the early reflections are also of interest [18].

6. REFERENCES

[1] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[2] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[3] R. Kumaresan and D. W. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 19, no. 1, pp. 134–139, Jan. 1983.

[4] M. Viberg, B. Ottersten, and T. Kailath, "Detection and estimation in sensor arrays using weighted subspace fitting," *IEEE Trans. Signal Process.*, vol. 39, no. 11, pp. 2436–2449, Nov. 1991.

[5] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[6] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 5, pp. 1210–1217, Oct. 1983.

[7] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 823–831, Aug. 1985.

[8] K. M. Buckley and L. J. Griffiths, "Broad-band signal-subspace spatial-spectrum (BASS-ALE) estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 7, pp. 953–964, Jul. 1988.

[9] C. H Knapp and G. C. 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.

[10] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Comput. Speech Language*, vol. 9, no. 2, pp. 153–169, Apr. 1995.

[11] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[12] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, Dec. 1987.

[13] M. Wohlmayr and M. Képesi, "Joint position-pitch extraction from multichannel audio," in *Proc. Interspeech*, Aug. 2007, pp. 1629–1632.

[14] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923–933, May 2013.

[15] J. R. Jensen, M. G. Christensen, J. Benesty, and S. H. Jensen, "Joint spatio-temporal filtering methods for DOA and fundamental frequency estimation," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 23, no. 1, pp. 174–185, Jan. 2015.

[16] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown University, May 2000.

[17] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 281–295, Feb. 1996.

[18] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. National Academy of Sciences*, vol. 110, no. 30, pp. 12186–12191, 2013.

[19] J. R. Jensen, J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "On frequency domain models for TDOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 11–15.

[20] J. Li and R. Wu, "An efficient algorithm for time delay estimation," *IEEE Trans. Signal Process.*, vol. 46, no. 8, pp. 2231–2235, Aug. 1998.

[21] Y. Wang, J. Li, P. Stoica, M. Sheplak, and T. Nishida, "Wideband RELAX and wideband CLEAN for aeroacoustic imaging," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 757–767, Feb. 2004.

[22] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Classics in Applied Mathematics. SIAM, 1995.

[23] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2010, Ver. 2.0.20100920.

[24] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.