NEAR-FIELD LOCALIZATION OF AUDIO: A MAXIMUM LIKELIHOOD APPROACH

Jesper Rindom Jensen and Mads Græsbøll Christensen

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

ABSTRACT

Localization of audio sources using microphone arrays has been an important research problem for more than two decades. Many traditional methods for solving the problem are based on a two-stage procedure: first, information about the audio source, such as time differences-of-arrival (TDOAs) and gain ratios-of-arrival (GROAs) between microphones is estimated, and, second, this knowledge is used to localize the audio source. These methods often have a low computational complexity, but this comes at the cost of a limited estimation accuracy. Therefore, we propose a new localization approach, where the desired signal is modeled using TDOAs and GROAs, which are determined by the source location. This facilitates the derivation of one-stage, maximum likelihood methods under a white Gaussian noise assumption that is applicable in both near- and far-field scenarios. Simulations show that the proposed method is statistically efficient and outperforms state-of-the-art estimators in most scenarios, involving both synthetic and real data.

Index Terms— Audio localization, microphone array, maximum likelihood, near-field, time difference-of-arrival, gain ratio-of-arrival.

1. INTRODUCTION

Recent technological advances have made microphones cheaper and smaller, and, as an effect of that, multiple microphones are nowadays integrated into many electrical devices such as televisions, smart phones, hearing aids, smart homes, etc. This evolution is particularly interesting from a signal processing perspective as multichannel recordings facilitate automatic camera steering, beamforming, dereverberation, and surveillance. Naturally, this has spawned much interest in developing new methods for localization of audio sources, i.e., methods for finding the position of an audio source in relation to an array of microphones.

While localization of audio sources is even more important now with the current technology, some of the pioneering work on audio localization dates back to the early 1980s. The methods proposed for localization can roughly be divided into two categories, i.e., those for estimating the direction-of-arrival (DOA) of the source (e.g., [1]), and those for estimating the position (i.e., DOA and range) of the source in relation to a microphone array (e.g., [2]). The latter group can generally be used no matter if the source is in the near- or far-field of the array, since the range information will be available, and can be exploited to account for near-field phenomena. In this paper, we therefore restrict ourselves to consider the topic of localization in the form of position estimation (denoted as localization in the remainder of the paper), since the objective is near-field localization of audio.

Far most localization methods for audio sources have traditionally been based on a two-stage approach. In these, a set of time differences-of-arrival (TDOAs) of the source between the different microphones are estimated first, whereupon these are utilized to obtain an estimate of the location of the source. Some classical, and still widely used, methods for TDOA estimation were proposed in [3]. Later, it was considered in, e.g., [4], and some of the references therein, how the localization is then performed given the TDOA information. Other approaches [5, 6] considered how gain ratios-of-arrival (GROAs) between microphones can be exploited for localization due to the inverse square law for sound radiation, and even how TDOAs and GROAs can be used jointly for localization [7]. As noted in [4], the two-stage approach often results in computationally fast localization algorithms but, unfortunately, at the cost of lower estimation accuracy compared to a single-stage approach.

In this paper, we take a different, single-stage approach to localization. First, we model the multichannel audio signal using both TDOAs and GROAs, and we assume that the audio source is periodic, which is reasonable for short segments of voiced speech and many musical instrument signals [8]. The TDOAs and GROAs are then further modeled using the source-to-array center distance (SAD) of microphones and the DOA. Based on this model, and under a white Gaussian noise assumption, we then derive different (asymptotic) maximum likelihood estimators of the SAD and DOA, two parameters revealing the source location.

The remainder of the paper is organized as follows. In Section 2, we introduce the signal model and find the likelihood function of the observed signal. We then show how the likelihood function can be used for SAD and DOA estimation through maximization in Section 3. The proposed estimators are then evaluated in Section 4, and, finally, the paper concludes with a discussion in Section 5, relating the work to state of the art.

This research was funded by the Danish Council for Independent Research, grant ID: DFF 1337-00084, and the Villum Foundation.

2. SIGNAL MODEL AND LIKELIHOOD

Consider a scenario where K microphones in an enclosure are each utilized to acquire N time-consecutive data snapshots, $x_k(n)$, for $n = 0, \dots, N-1$, where n denotes the observation time index and k the microphone number for $k = 1, \dots, K$. These observations can be stacked in to observation vectors $\mathbf{x}_k \in \mathbb{C}^N$ as $\mathbf{x}_k = \begin{bmatrix} x_k(0) & x_k(1) & \cdots & x_k(N-1) \end{bmatrix}^T$. We then assume that a periodic source of interest and sensor noise is present in the recording environment, and that the recording environment is anechoic. Since the sensors are located with different distances to the source, and due to the inverse square law for sound radiation, the observations at each sensor are a sum of a delayed and attenuated version of a periodic signal, $s(n)$, and noise, $v_k(n)$. The observed signal at microphone k at time instance n is, therefore, given by

$$x_k(n) = \beta_k s(n - f_\text{s}\tau_k) + v_k(n), \tag{1}$$

where β_k is the attenuation of the source from its position to microphone k, f_s is the sampling frequency, and τ_k is the time it takes the source to travel to microphone k. Let us then choose sensor one as our reference such that $s_1(n) = \beta_1 s(n - f_\text{s}\tau_1)$. With the choice of reference, the observed signal model can be rewritten as

$$\begin{aligned} x_k(n) &= \frac{r_1}{r_k} s_1(n - f_\text{s}\tau_{1k}) + v_k(n) & (2)\\ &= \frac{r_1}{r_k} s_1\left(n - f_\text{s}\frac{r_k - r_1}{c}\right) + v_k(n), & (3) \end{aligned}$$

with r_k being the distance from microphone k to the source, and c is the wave propagation speed. If we then assume that the microphones are organized in a known array structure, we can further model r_k. In the remainder of the paper, we assume a uniform linear array (ULA) structure, but the derivations herein can easily be modified to other array structures. In the ULA case, it can be shown using the law of cosines that the distance r_k is given by $r_k = \sqrt{g_k^2 d^2 + r_\text{c}^2 - 2g_k d r_\text{c} \sin\theta}$, where $g_k = \frac{K-1}{2} - k + 1$, d is the spacing between the microphones, r_c is the SAD, and θ is the DOA of the source onto the array. Furthermore, since we assume $s_1(n)$ is a periodic signal, we can model it as

$$s_1(n - f_\text{s}\frac{r_k - r_1}{c}) = \sum_{l=1}^{L} \gamma_l e^{jl\omega_0 n} e^{-j2\pi l f_0 \frac{r_k - r_1}{c}} \tag{4}$$

where L is the model order, i.e., the number of harmonics constituting the periodic signal, ω_0 is the fundamental frequency, $\gamma_l = \beta_1 \alpha_l$, $\alpha_l = A_l e^{\phi_l}$ is the complex amplitude of the lth harmonic, A_l is its real amplitude, ϕ_l is its phase, and $f_0 = f_\text{s}\frac{\omega_0}{2\pi}$. When the model holds and is exploited by the estimator, we can potentially get more robust and accurate estimates compared to methods integrating over broad frequency ranges as discussed in [9]. While the above model is for complex signals, it can be applied on real data by using the Hilbert transform. Note that, in this work, we consider the fundamental frequency as a known parameter, and in practice it can be estimated using the statistically efficient, multichannel, pitch estimator in [10]. Using the aforementioned models for the distances and the periodic signal, we can rewrite the model of the observed signal as

$$x_k(n) = \frac{r_1}{r_k} \sum_{l=1}^{L} \gamma_l e^{jl\omega_0 n} e^{-j2\pi l f_0 \frac{r_k - r_1}{c}} + v_k(n). \tag{5}$$

Eventually, the model for $x_k(n)$ can be used to model the observed signal vector as

$$\mathbf{x}_k = \mathbf{Z}(\omega_0)\mathbf{D}_k(r_\text{c}, \theta)\boldsymbol{\gamma} + \mathbf{v}_k, \tag{6}$$

with $\boldsymbol{\gamma} = [\gamma_1 \ \cdots \ \gamma_L]^T$, $\mathbf{Z}(\omega_0) = [\mathbf{z}(\omega_0) \ \cdots \ \mathbf{z}(L\omega_0)]$, $\mathbf{z}(\omega) = [1 \ e^{j\omega} \ \cdots \ e^{j(N-1)\omega}]^T$, $\mathbf{v}_k = [v_k(0) \ \cdots \ v_k(N-1)]^T$,

$$[\mathbf{D}_k(r_\text{c}, \theta)]_{ll} = \sqrt{\frac{g_1^2 d^2 + r_\text{c}^2 - 2g_1 d r_\text{c} \sin\theta}{g_k^2 d^2 + r_\text{c}^2 - 2g_k d r_\text{c} \sin\theta}} e^{-j2\pi l f_0 \frac{w_k(r_\text{c}, \theta)}{c}}, \tag{7}$$

and $[\mathbf{D}_k(r_\text{c}, \theta)]_{pq} = 0$ for $p \neq q$, where $[\cdot]_{pq}$ denotes the (p, q)'th element of a matrix, and

$$\begin{aligned} w_k(r_\text{c}, \theta) = &\sqrt{g_k^2 d^2 + r_\text{c}^2 - 2g_k d r_\text{c} \sin\theta} \\ &- \sqrt{g_1^2 d^2 + r_\text{c}^2 - 2g_1 d r_\text{c} \sin\theta}. \end{aligned} \tag{8}$$

If we assume that the noise is white Gaussian in each channel, and that the noise is uncorrelated across channels, it can be shown that the log-likelihood function for our set of observation vectors is given by [10, 11]

$$\begin{aligned} &\ln p(\{\mathbf{x}_k(n)\}; \boldsymbol{\psi}) = \\ &\quad -NK \ln \pi - N \sum_{k=1}^{K} \ln \sigma_k^2 - \sum_{k=1}^{K} \frac{\|\mathbf{v}_k(n)\|^2}{\sigma_k^2}, \end{aligned} \tag{9}$$

where $\boldsymbol{\psi}$ is a vector containing the unknown signal parameters of interest, and σ_k^2 is the variance of the noise at microphone k. The goal is then to estimate r_c and θ given the set of observed signal vectors $\{\mathbf{x}_k(n)\}_{k=1}^{K}$, as these parameters reveal the location of the periodic source relative to the array center.

3. LOCALIZATION METHODS

We then proceed with deriving optimal source DOA and SAD by maximizing the log-likelihood function in (9). First, the log-likelihood is maximized with respect to the unknown amplitudes $\boldsymbol{\gamma}$. Differentiating (9) with respect $\boldsymbol{\gamma}$, equating with zero, and solving for the unknown amplitudes, yields the following estimates:

$$\widehat{\boldsymbol{\gamma}} = \left(\sum_{k=1}^{K} \frac{\mathbf{D}_k^H \mathbf{Z}^H \mathbf{Z} \mathbf{D}_k}{\sigma_k^2}\right)^{-1} \sum_{k=1}^{K} \frac{\mathbf{D}_k^H \mathbf{Z}^H \mathbf{x}_k}{\sigma_k^2}. \tag{10}$$

In a similar way, we solve for the unknown noise variance, which yields

$$\widehat{\sigma}_k^2 = N^{-1}\|\mathbf{x}_k - \mathbf{Z}\mathbf{D}_k\boldsymbol{\gamma}\|^2. \tag{11}$$

Note that the amplitude estimates depend on the noise variance and vice versa. In practice, we can deal with this issue by estimating the parameters iteratively and by initializing, e.g., the noise variances as $\sigma_k^2 = 1$ for $k = 1, \dots, K$. From our simulations, we experienced 2–3 iterations to be sufficient for achieving convergence in most scenarios.

After convergence, we can insert the noise variance estimate in (11) into (9). Obviously, the DOA and SAD can then be estimated by minimizing the sum of the logarithms of the noise variance estimates for the different microphones, i.e.,

$$\{\widehat{\theta}, \widehat{r}_\mathrm{c}\} = \arg \min_{\{\theta, r_\mathrm{c}\} \in \Theta \times \mathcal{R}_\mathrm{c}} \sum_{k=1}^{K} \ln \|\mathbf{x}_k - \mathbf{Z}\mathbf{D}_k\widehat{\boldsymbol{\gamma}}\|^2, \tag{12}$$

where Θ and $\mathcal{R}_\mathrm{c}$ are sets of candidate DOAs and SADs, respectively. In the remainder of the paper, we denote this estimator as the amplitude- and phase-based NLS (NLS-AP) estimator. By introducing different simplifications and assumptions, we can obtain computationally simpler algorithms as shown in the remainder of the section.

Simplification no. 1: We also derive the joint DOA and SAD estimator, using phase differences only. This simpler estimator can potentially yield better estimates when, e.g., the microphones have different gains if this is not accounted for. The corresponding observed signal model is:

$$\mathbf{x}_k = \beta_k \mathbf{Z}\mathbf{D}'_k\boldsymbol{\alpha} + \mathbf{v}_k, \tag{13}$$

where $[\mathbf{D}'_k]_{ll} = e^{-j2\pi l f_0 \frac{w_k(r_\mathrm{c},\theta)}{c}}$, $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \cdots & \alpha_L\end{bmatrix}^T$, and $[\mathbf{D}'_k]_{pq} = 0$ for $p \neq q$. The log-likelihood of the observations are equal to the one in (9), and the unknown β_k's, σ_k^2's, and $\boldsymbol{\alpha}$ can be found by using the following equations iteratively:

$$\widehat{\boldsymbol{\alpha}} = \left(\sum_{k=1}^{K} \frac{\beta_k^2}{\sigma_k^2}\mathbf{D}_k'^H\mathbf{Z}^H\mathbf{Z}\mathbf{D}'_k\right)^{-1} \sum_{k=1}^{K}\frac{\beta_k}{\sigma_k^2}\mathbf{D}_k'^H\mathbf{Z}^H\mathbf{x}_k, \tag{14}$$

$$\widehat{\beta}_k = \frac{\mathrm{Re}\{\boldsymbol{\alpha}^H\mathbf{D}_k'^H\mathbf{Z}^H\mathbf{x}_k\}}{\boldsymbol{\alpha}^H\mathbf{D}_k'^H\mathbf{Z}^H\mathbf{Z}\mathbf{D}'_k\boldsymbol{\alpha}}, \ \widehat{\sigma}_k^2 = \frac{\|\mathbf{x}_k - \beta_k\mathbf{Z}\mathbf{D}'_k\boldsymbol{\alpha}\|^2}{N}. \tag{15}$$

The iterative procedure can be initialized with, e.g., $\beta_k = 1$ and $\sigma_k^2 = 1$ for $k = 1, \dots, K$. After convergence, the DOA and SAD can then be estimated jointly by solving

$$\{\widehat{\theta}, \widehat{r}_\mathrm{c}\} = \arg \min_{\{\theta, r_\mathrm{c}\} \in \Theta \times \mathcal{R}_\mathrm{c}} \sum_{k=1}^{K} \ln \|\mathbf{x}_k - \widehat{\beta}_k\mathbf{Z}\mathbf{D}'_k\widehat{\boldsymbol{\alpha}}\|^2. \tag{16}$$

This estimator is denoted the phase-based NLS (NLS-P) estimator.

Simplification no. 2: Another possibility is to use only the information about the attenuations across the microphones to estimate the DOA and SAD, which can be advantageous when the microphones are not synchronized properly. If we are only using amplitude information, we can model our observations as

$$\mathbf{x}_k = \mathbf{Z}\boldsymbol{\gamma}_k + \mathbf{v}_k, \tag{17}$$

where $\boldsymbol{\gamma}_k = \begin{bmatrix}\gamma_{k,1} & \cdots & \gamma_{k,L}\end{bmatrix}^T$, $\gamma_{k,l} = \beta_k\alpha_{k,l}$, $\alpha_{k,l} = A_l e^{j(\phi_l + \Delta_{k,l})}$, and $\Delta_{k,l}$ is a phase shift of the l'th harmonic arising from the travel time from the source to microphone k. The amplitudes, $\boldsymbol{\gamma}_k$, can be estimated in a maximum likelihood sense, by maximizing the aforementioned likelihood function for the above signal model. This yields

$$\widehat{\boldsymbol{\gamma}}_k = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}_k. \tag{18}$$

From the inverse square law of sound radiation, we know that $\|\boldsymbol{\gamma}_p\| = \frac{r_q}{r_p}\|\boldsymbol{\gamma}_q\|$. That is, the estimated amplitudes, $\widehat{\boldsymbol{\gamma}}_k$, can be used to estimate the DOA and SAD by solving

$$\{\widehat{\theta}, \widehat{r}_\mathrm{c}\} = \arg \min_{\{\theta, r_\mathrm{c}\} \in \Theta \times \mathcal{R}_\mathrm{c}} \sum_{p=1}^{K} \sum_{\substack{q=1,\\ q\neq p}}^{K} \left(\|\gamma_p\| - \frac{r_p}{r_q}\|\gamma_q\|\right)^2, \tag{19}$$

denoted as the amplitude-based NLS (NLS-A) estimator.

In summary, we proposed to estimate the DOA and SAD of a broadband audio source in relation to a microphone array by exploiting the structure of the source, which is here assumed to be harmonic. In the more general case, where this assumption does not hold, we can estimate the DOA and SAD by setting $L = 1$, and integrate the likelihood over range of frequency bins for sets of candidate DOAs and SADs, and then maximize the integrated likelihoods.

4. EXPERIMENTAL RESULTS

First, the proposed methods were evaluated on synthetic signals. In these evaluations, a synthetic harmonic signal with 4 unit amplitude harmonics was generated and added to white Gaussian noise. The methods proposed herein (NLS-AP, NLS-P, and NLS-A), and a couple of reference methods (the SRP-PHAT [12] and WLSWM [11] methods) were then applied on this data for estimation of the DOA and the distance to the source. Note that the fundamental frequency is needed in the proposed methods, and it was therefore estimated using the ML multichannel pitch estimator in [10]. For each generated synthetic signal, 100 Monte-Carlo simulations were conducted in this way, and the noise and the phases of the harmonics was randomized in each simulation. The mean squared errors (MSEs) of the parameter estimates were measured across the Monte-Carlo simulations for each setting. In the first experiment, we evaluated the MSEs of the methods for different signal-to-noise ratios (SNRs), with the following choice of other parameters: $N = 30$, $K = 3$, $c = 343$ m/s, $d = 0.05$, $f_0 = 263$ Hz, $f_\mathrm{s} = 4$ kHz, $r_\mathrm{c} = 0.5$ m and

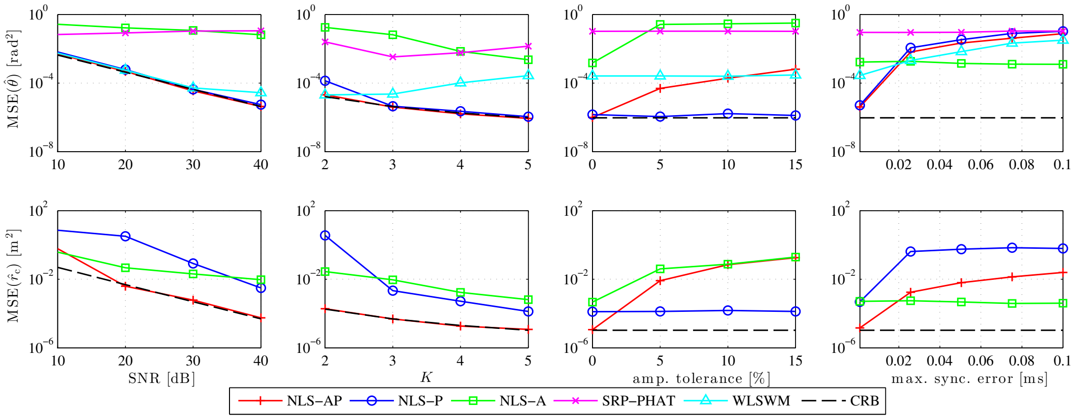

Fig. 1. MSEs of the DOA and SAD estimates obtained using the proposed methods, and the SRP-PHAT and WLSWM methods versus (from left to right) the SNR, K, θ, and r_c.

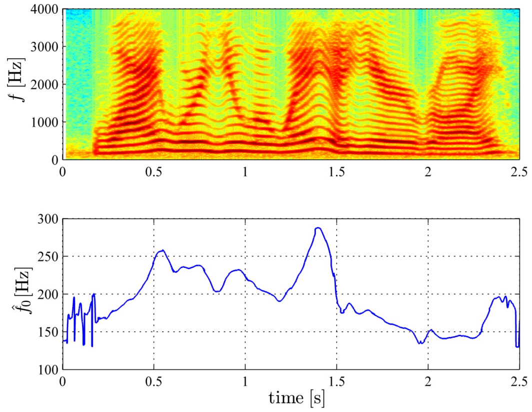

Fig. 2. Plots of (top) the spectrogram of a speech signal, and (bottom) the estimated fundamental frequency track.

$\theta = 45°$. Moreover, in this and all the following simulations on synthetic data, the cost-function in the SRP-PHAT method is obtained by integrating over the frequency interval $[200; f_s/2]$ Hz, and the FFT length was 256. Next, the MSEs were measured for different K's. For this experiment, SNR $= 40$ dB while the other parameters were the same. In the two final experiments on synthetic data, the performances of the different methods were measured in two cases of uncertainties: 1) when each microphone had a gain tolerance, i.e., each amplitude could deviate from the model in (5) by some percentage; and 2) when the microphones are not perfectly synchronized, i.e., an extra, uniformly distributed, random delay was added to each channel. In these simulations, we had SNR $= 40$ dB, $N = 30$, $K = 5$, and the remaining parameters was set as in the previous simulations. All the results are shown in Fig. 1, where they are also compare with the Cramér-Rao bound (CRB). Regarding DOA estimation, we observe that the proposed NLS-AP and NLS-P estimators show similar performance in most cases without uncertainties, and that they are statistically efficient. For low SNRs and K's, the performance of the WLSWM method is comparable to that of the aforementioned proposed methods. Otherwise, the proposed NLS-AP and NLS-P methods clearly outperform the other methods. Regarding SAD estimation with no uncertainties, the proposed NLS-AP method is statistically efficient and outperforms the other proposed methods, except for low SNR's where the NLS-A method shows similar performance. When we have uncertainties, the results are quite different. Expectedly, the NLS-P shows superior performance compared to all other methods when there is a tolerance on the amplitudes. This happens already when the tolerance exceeeds $\approx$ 2 %. Furthermore, the NLS-A has the better performance compared to all the other methods, when we have synchronization errors that can be larger than 0.02 ms.

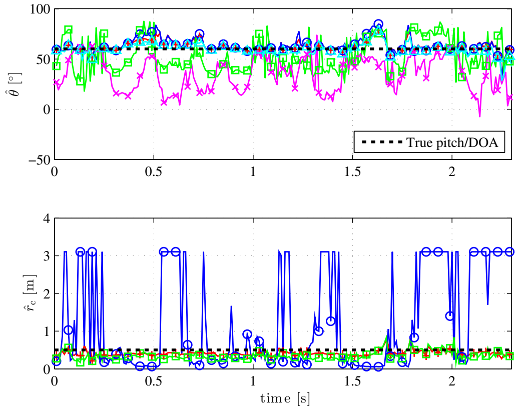

Fig. 3. The estimates obtained with different DOA and SAD estimators when applied on the speech signal for $T_{60} = 0.1$ s.

The same methods were also evaluted on a speech signal with the spectrogram and pitch track shown in Fig. 2. The pitch track was estimated using the approximate NLS method proposed in [8]. The utilized signal was single-channel, and, therefore, resynthesized spatially using the online available room impulse response (RIR) generator [13]. The RIR generator was set up as follows: $c = 343$ m/s, $f_s = 8$ kHz, the microphones of a ULA was located at $[2 + d(k - \frac{K-1}{2})]$ m $\times$ 0.1 m $\times$ 1.5 m for $k = 1, \ldots, K$, $d = 0.05$ m, the source was located at $\theta = 60°$ and $r_c = 0.5$, the room dimensions was 4 m $\times$ 4 m $\times$ 3 m, the length of the RIRs was 2,048, the microphone type was cardioid, and they where oriented with an azimuth of 90° and an elevation of 0°. With this setup, we generated the spatial-temporal data on which the aforementioned methods were applied on consecutive frames of length $N = 50$ of the multichannel signal with $K = 3$, and in the SRP-PHAT method we integrated over the frequencies in the interval $[150, f_\mathrm{s}/2]$ Hz. This experiment was conducted for a reverberation time of $T_{60} = 0.1$ s. The resulting estimates are depicted in Fig. 3. We can see that the NLS-AP, NLS-P, and WLSWM seem to provide the most accurate DOA estimates followed by the NLS-A and SRP-PHAT methods. These results are interesting, as they indicate that the proposed methods are robust against small degrees of reverberation although this is not explicitly accounted for in the model. In terms of SAD estimation, the NLS-AP and NLS-A method clearly outperforms the NLS-P method.

5. DISCUSSION

In this paper, localization of audio sources using a microphone array have been considered. Through the past couple of decades, many methods [4–7] have been proposed for solving this research problem, with many of those being based on a two-stage procedure. First, they estimate, e.g., either the TDOAs, the GROAs, or both of the audio source between the different microphones in the array. Then, they use these parameter estimates to form an estimate of the location of the audio source. As stated in [4], these two-stage procedures typically have a low computational complexity, but at the cost of a limited estimation accuracy. Herein, we therefore propose a new, single-stage approach, where a the desired signal is modeled using TDOAs and GROAs, and by assuming that the desired signal is periodic. With this model, we show how maximum likelihood estimates of the audio source location can be obtained, when the noise is white Gaussian. That is, the proposed method is statistically efficient in near-field as well as far-field scenarios, which, to our knowledge, is not the case for any other method except for the one in [7]. However, the method in [7] requires knowledge about TDOAs and GROAs as opposed to the proposed methods. While the proposed methods are derived for periodic signals, it should be noted that they can be applied on general, broadband sources by setting $L = 1$, and integrating the likelihood over a range of frequency bins for sets of candidate DOAs and SADs, and by maximizing the integrated likelihoods. The simulation results show statistical efficiency of the proposed method, and shows that it outperforms the widely used SRP-PHAT method [12] as well as an efficient far-field DOA estimation method proposed in [11]. Moreover, simulations show that the proposed method is relatively robust against reverberance even though this is not accounted for in the derivations.

REFERENCES

[1] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, “Localization of multiple acoustic sources with small arrays using a coherence test,” *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. 2136–2147, Apr. 2008.

[2] P. Bestagini, M. Compagnoni, F. Antonacci, A. Sarti, and S. Tubaro, “TDOA-based acoustic source localization in the space–range reference frame,” *Multidim. Syst. Sign. Process.*, pp. 1–23, Mar. 2013.

[3] 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.

[4] M. D. Gillette and H. F. Silverman, “A linear closed-form algorithm for source localization from time-differences of arrival,” *IEEE Signal Process. Lett.*, vol. 15, pp. 1–4, Jan. 2008.

[5] X. Sheng and Y.-H. Hu, “Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks,” *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 44–53, Jan. 2005.

[6] S. T. Birchfield and R. Gangishetty, “Acoustic localization by interaural level difference,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 4, pp. 1109–1112.

[7] K. C. Ho and M. Sun, “Passive source localization using time differences of arrival and gain ratios of arrival,” *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 464–477, Feb. 2008.

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

[9] 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.

[10] M. G. Christensen, “Multi-channel maximum likelihood pitch estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.

[11] J. R. Jensen, M. G. Christensen, and S. H. Jensen, “Statistically efficient methods for pitch and DOA estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013.

[12] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., chapter 8, pp. 157–180. Springer-Verlag, 2001.

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