# Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation

Jensen, Jesper Rindom; Christensen, Mads Græsbøll; Jensen, Søren Holdt

*Published in:*
I E E E Transactions on Audio, Speech and Language Processing

*DOI (link to publication from Publisher):*
10.1109/TASL.2013.2239290

*Publication date:*
2013

*Document Version*
Accepted author manuscript, peer reviewed version

# Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation

Jesper Rindom Jensen, *Member, IEEE,* Mads Græsbøll Christensen, *Senior Member, IEEE,* and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*—In this paper, we consider the problem of joint direction-of-arrival (DOA) and fundamental frequency estimation. Joint estimation enables robust estimation of these parameters in multi-source scenarios where separate estimators may fail. First, we derive the exact and asymptotic Cramér-Rao bounds for the joint estimation problem. Then, we propose a nonlinear least squares (NLS) and an approximate NLS (aNLS) estimator for joint DOA and fundamental frequency estimation. The proposed estimators are maximum likelihood estimators when: 1) the noise is white Gaussian, 2) the environment is anechoic, and 3) the source of interest is in the far-field. Otherwise, the methods still approximately yield maximum likelihood estimates. Simulations on synthetic data show that the proposed methods have similar or better performance than state-of-the-art methods for DOA and fundamental frequency estimation. Moreover, simulations on real-life data indicate that the NLS and aNLS methods are applicable even when reverberation is present and the noise is not white Gaussian.**

***Index Terms*—Direction-of-arrival estimation, fundamental frequency estimation, joint estimation, non-linear least squares, Cramér-Rao lower bound.**

## I. Introduction

BOTH direction-of-arrival (DOA) estimation and fundamental frequency estimation are very important signal processing topics, and, individually, these two estimation problems are widely studied research topics. DOA estimation, for example, has been treated in many text books and research papers (see, e.g., [1]–[6]) and has a multitude of applications in areas such as geophysics, radio astronomy, biomedical engineering, radar and microphone arrays. Fundamental frequency estimation (we will also refer to this as pitch estimation), on the other hand, has applications such as compression, separation and enhancement of, e.g., audio and voiced speech [7], [8], automatic music transcription and music classification [9]. For an overview of existing pitch estimation techniques, see, e.g., [9]–[13]. That is, both DOA and pitch estimation are relevant for processing of audio and speech signals. A few examples of applications which can benefit from the knowledge of both the DOA and the pitch are hands-free communication, teleconferencing, surveillance applications and hearing aids.

It is therefore natural to consider joint spatio-temporal processing of audio and speech signals which is the topic of this paper. More specifically, we consider joint DOA and pitch estimation. Besides the convenience of being able to estimate the DOA and the pitch simultaneously, joint spatio-temporal processing potentially has two significant advantages. For instance, if a source parameter is similar for both sources in a two-source scenario, the sources are not resolvable if we only estimate this parameter separately. If joint parameter estimation of several parameters is performed and just some of the parameters are distinct, then the sources are possibly still resolvable. Another important advantage of joint estimation relates to the estimation accuracy. For example, DOA and pitch estimation of periodic sources such as many musical instruments and voiced speech can be conducted by first estimating the DOA, then by extracting the signal impinging from that DOA, and, finally, by estimating the pitch from the extracted signal. However, the extraction can be seen as a linear data transformation which most likely increases the Cramér-Rao bound (CRB) for the pitch estimate, meaning that the resulting estimates may be suboptimal. Other important issues regarding processing of multi-channel signals are, e.g., reverberation and array calibration errors. We refer the interested reader to [14], [15] for an overview of methods dealing with these problems as these topics are out of the scope of this paper.

Motivated by the above observations, and due to an increasing computational capability, the computationally demanding problem of joint DOA and pitch estimation has attracted considerable attention in the recent years. As a result, some methods have been proposed for solving the joint estimation problem. Basically, these methods can be divided into two groups. The first group jointly estimates the frequency and the DOA of a single sinusoid defined in two dimensions (e.g., time and space). A few examples of such methods are [16], where a state-space realization technique is used, [17]–[19] which is based on the 2-D minimum variance distortionless response (MVDR) method, [20] which is based on the ESPRIT method, and [21] where a signal-dependent multistage Wiener filter (MWF) [22] is used. This group of methods is not commonly used in speech and audio processing. In most of the literature (see, e.g., [23]–[27]), DOA estimation of audio and speech recorded using a microphone array, has been treated as a broadband problem. In this paper, however, we shall assume a harmonic model which describes, e.g., many musical instruments and voiced speech well; this will incontrovertibly

Manuscript received March 03, 2012; revised June 21, 2012; accepted January 02, 2013. 

This work was supported in part by the Villum foundation.

J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, Dept. of Arch., Design & Media Technology, Aalborg University, 9200 Aalborg, Denmark, e-mails: {jrj,mgc}@create.aau.dk.

S. H. Jensen is with the Dept. of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark, e-mail: shj@es.aau.dk.

Digital Object Identifier XXXX

enable us to treat the joint DOA and pitch estimation problem as $L$ narrowband problems. Methods utilizing this fact forms the other group of estimators that consider the case with one or more harmonically related, two-dimensional sinusoids. These methods can, therefore, be seen as a generalization of the first group of methods. A few methods dealing with this case have been proposed; in [28] a ML-based method is proposed; in [29]–[31] subspace-based methods are proposed; in [32], [33] a correlation-based method is proposed; and in [34], [35] some spatio-temporal filtering methods based on the linearly constrained minimum variance (LCMV) beamformer [36] and the periodogram are proposed. Note that some of the above-mentioned methods consider time delay estimation and not DOA estimation, however, these two parameters are closely related.

In this paper, we also consider joint DOA and pitch estimation. Based on the harmonic model, we derive the exact and asymptotic CRBs for the joint DOA and pitch estimation problem. Moreover, we propose a non-linear least squares (NLS) method for joint DOA and pitch estimation. The proposed estimator is derived under the assumptions that the noise is white Gaussian, the array is a uniform linear array, the environment is anechoic, and the source of interest is located in the far-field of the array. When the assumptions hold, the proposed NLS estimator is also the maximum likelihood (ML) estimator as opposed to most of the existing joint DOA and pitch estimators [28]–[35]. Moreover, the proposed estimator is applicable in scenarios with any number of sensors, and it is easily generalized to support any array structure as opposed to the joint estimators in [29], [30]. Finally, we propose an approximate NLS (aNLS) method which is computationally more efficient.

The rest of the paper is organized as follows: in Section II, we introduce the spatio-temporal harmonic signal model. Then, in Section III, we derive the exact and asymptotic CRBs for the joint DOA and pitch estimation problem; the asymptotic bounds are used to motivate why the DOA and pitch should be estimated jointly. We derive the NLS and aNLS estimators for joint DOA and pitch estimation in Section IV, and the estimators are evaluated on synthetic as well as real-life signals in Section V. Finally, Section VI concludes our work.

## II. Signal Model

In this paper, we consider joint estimation of the DOA, $\theta$, and the pitch, $\omega_0$, of a quasi-periodic source, also referred to as the source of interest (SOI), which is recorded using a $N_\mathrm{s}$-element uniform linear array (ULA) in a noisy and anechoic environment. We assume that the noise is uncorrelated with the SOI. The ULA and the definition of $\theta$ are illustrated in Fig. 1. Real-life examples of quasi-periodic sources are, e.g., voiced speech and musical instruments. We assume that the quasi-periodic source is in the far-field of the ULA. The signal measured on the $n_\mathrm{s}$th sensor at time instance $n_\mathrm{t}$ for $n_\mathrm{s} = 0, \ldots, N_\mathrm{s}-1$ and $n_\mathrm{t} = 0, \ldots, N_\mathrm{t}-1$ is then given by

$$\begin{aligned} y_{n_\mathrm{s}}(n_\mathrm{t}) &= \beta_{n_\mathrm{s}} s(n_\mathrm{t} - f_\mathrm{s}\tau_{n_\mathrm{s}}) + w_{n_\mathrm{s}}(n_\mathrm{t}) \\ &= x_{n_\mathrm{s}}(n_\mathrm{t}) + w_{n_\mathrm{s}}(n_\mathrm{t}) \ , \end{aligned} \tag{1}$$

where $\beta_{n_\mathrm{s}}$ and $\tau_{n_\mathrm{s}}$ are the attenuation and the delay of the wave generated by the SOI from sensor 0 to sensor $n_\mathrm{s}$, respectively, $f_\mathrm{s}$ is the sampling frequency, $s(n_\mathrm{t} - f_\mathrm{s}\tau_{n_\mathrm{s}})$ is the delayed quasi-periodic signal, and $w_{n_\mathrm{s}}(n_\mathrm{t})$ is the noise picked up by the $n_\mathrm{s}$th sensor. Note that, in the rest of the paper, $(\cdot)_{n_\mathrm{s}}$ means that the variable or constant is related to the $n_\mathrm{s}$th sensor. Due to the array structure, we know that the delay is given by

$$\tau_{n_\mathrm{s}} = n_\mathrm{s}\frac{d\sin\theta}{c} \ , \tag{2}$$

where $d$ is the inter-element spacing of the ULA, and $c$ is the wave propagation velocity. Since the SOI is assumed to be quasi-periodic, we know that it can be modeled as a harmonic source

$$s(n_\mathrm{t}) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n_\mathrm{t}} \ , \tag{3}$$

for $n_\mathrm{t} = 0, \ldots, N_\mathrm{t}-1$, where $L$ is the model order, $\alpha_l = A_l e^{j\phi_l}$, and $A_l > 0$ and $\phi_l$ are the real amplitude and phase of the $l$th harmonic. In case the desired signal has inharmonicities, the model can be extended to account for this [12], [37], [38]. Note that the signal model is complex as opposed to many real-life signals which are real. However, it is common to use complex signal representations since it leads to a simpler notation, and the complex model can easily be applied on real signals if we convert these to analytic signals using the Hilbert transform [6], [9]. In this paper, we consider the model order $L$ as a known parameter (see, e.g., [6], [39] and the references therein for an overview of existing model order estimators).

Using the signal model in (3), the desired signal at sensor $n_\mathrm{s}$ can be written as

$$s(n_\mathrm{t} - f_\mathrm{s}\tau_{n_\mathrm{s}}) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0(n_\mathrm{t} - f_\mathrm{s}\tau_{n_\mathrm{s}})} \tag{4}$$

$$= \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n_\mathrm{t}} e^{-jl\omega_\mathrm{s} n_\mathrm{s}} \ , \tag{5}$$

where $\omega_\mathrm{s} = \omega_0 f_\mathrm{s}\tau_1$ is the so-called spatial frequency. Note that the spatial frequency is dependent on the fundamental frequency, $\omega_0$.

Additionally, we define a spatio-temporal matrix signal model, which is useful in the derivation of parameter estimators. The matrix model is defined as

$$\mathbf{Y}(n_\mathrm{t}) = \mathbf{X}(n_\mathrm{t}) + \mathbf{W}(n_\mathrm{t}) \ , \tag{6}$$

where

$$\mathbf{Y}(n_\mathrm{t}) = \begin{bmatrix} y_0(n_\mathrm{t}) & \cdots & y_0(n_\mathrm{t} - N_\mathrm{t} + 1) \\ \vdots & \ddots & \vdots \\ y_{N_\mathrm{s}-1}(n_\mathrm{t}) & \cdots & y_{N_\mathrm{s}-1}(n_\mathrm{t} - N_\mathrm{t} + 1) \end{bmatrix} \ , \tag{7}$$

with $\mathbf{X}(n_\mathrm{t})$ and $\mathbf{W}(n_\mathrm{t})$ being defined similarly to $\mathbf{Y}(n_\mathrm{t})$. The attenuated desired signal matrix $\mathbf{X}(n_\mathrm{t})$ can be rewritten using (5) as

$$\mathbf{X}(n_\mathrm{t}) = \boldsymbol{\beta} \sum_{l=1}^{L} \alpha_l(n_\mathrm{t}) \mathbf{z}_\mathrm{s}(l\omega_\mathrm{s}) \mathbf{z}_\mathrm{t}^T(l\omega_0) \ , \tag{8}$$

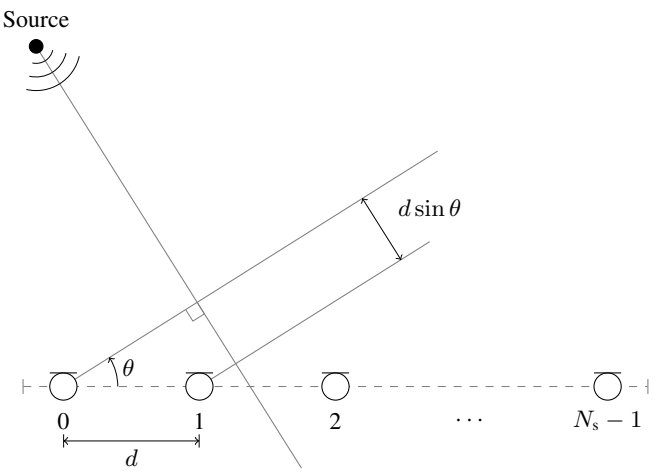


Fig. 1. Illustration of the uniform linear array structure assumed in this paper.

where

$$\alpha_l(n_\mathrm{t}) = \alpha_l e^{jl\omega_0 n_\mathrm{t}} \ , \tag{9}$$

$$\boldsymbol{\beta} = \mathrm{diag}\left\{\begin{bmatrix}\beta_0 & \cdots & \beta_{N_\mathrm{s}-1}\end{bmatrix}^T\right\} \ , \tag{10}$$

$$\mathbf{z}_\mathrm{s}(\omega_\mathrm{s}) = \begin{bmatrix}1 & e^{-j\omega_\mathrm{s}} & \cdots & e^{-j(N_\mathrm{s}-1)\omega_\mathrm{s}}\end{bmatrix}^T \ , \tag{11}$$

$$\mathbf{z}_\mathrm{t}(\omega_0) = \begin{bmatrix}1 & e^{-j\omega_0} & \cdots & e^{-j(N_\mathrm{t}-1)\omega_0}\end{bmatrix}^T \ , \tag{12}$$

with $\mathrm{diag}\{\cdot\}$ denoting the operator that transforms a vector into a diagonal matrix, and $(\cdot)^T$ denoting the transpose of a vector or matrix. Alternatively, the matrix model in (6) can be mapped to a vector model by stacking the columns of $\mathbf{Y}(n_\mathrm{t})$ as

$$\begin{aligned}\mathbf{y}(n_\mathrm{t}) &= \mathrm{vec}\{\mathbf{Y}(n_\mathrm{t})\} \\ &= \mathbf{x}(n_\mathrm{t}) + \mathbf{w}(n_\mathrm{t}) = \bar{\mathbf{Z}}\boldsymbol{\alpha}(n_\mathrm{t}) + \mathbf{w}(n_\mathrm{t}) \ ,\end{aligned} \tag{13}$$

where $\mathrm{vec}\{\cdot\}$ is the column-wise stacking operator, and

$$\bar{\mathbf{Z}} = \mathbf{B}\mathbf{Z} \ , \tag{14}$$

$$\mathbf{B} = \begin{bmatrix}\boldsymbol{\beta} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\beta}\end{bmatrix} \ , \tag{15}$$

$$\mathbf{Z} = \begin{bmatrix}\mathbf{z}(\omega_0,\omega_\mathrm{s}) & \cdots & \mathbf{z}(L\omega_0, L\omega_s)\end{bmatrix} \ , \tag{16}$$

$$\mathbf{z}(l\omega_0, l\omega_\mathrm{s}) = \mathbf{z}_\mathrm{t}(l\omega_0) \otimes \mathbf{z}_\mathrm{s}(l\omega_\mathrm{s}) \ , \tag{17}$$

$$\boldsymbol{\alpha}(n_\mathrm{t}) = \begin{bmatrix}\alpha_1 e^{j\omega_0 n_\mathrm{t}} & \cdots & \alpha_L e^{jL\omega_0 n_\mathrm{t}}\end{bmatrix}^T \ , \tag{18}$$

with $\otimes$ denoting the Kronecker product operator. In summary, the objective considered in this paper is to estimate the DOA and the pitch jointly from spatio-temporal, observed signal samples which can be modeled by (13).

## III. Cramér-Rao Bounds

It is common practice to place a lower bound on the variance of unbiased estimators. This is useful while evaluating the performance of such estimators, and it provides insight into the nature of the estimation problem. There exists a multitude of such bounds among which the CRB is one of the most commonly used [40]. In this section, we derive exact and asymptotic expressions for the CRBs for the joint DOA and pitch estimation problem. Moreover, we show why it is beneficial to estimate the DOA and the pitch jointly by analyzing the asymptotic CRB expressions.

### A. Exact Bounds

First, we derive the exact CRBs for the joint DOA and pitch estimation problem. Let

$$\bar{\mathbf{y}}(n_\mathrm{t}) = \begin{bmatrix}y_0(n_\mathrm{t}) & \cdots & y_{N_\mathrm{s}-1}(n_\mathrm{t})\end{bmatrix}^T \tag{19}$$

be the observed signal vector from the $N_\mathrm{s}$-element ULA at $n_\mathrm{t} \in [0; N_\mathrm{t}-1]$. We can also write the observation vector, $\bar{\mathbf{y}}(n_\mathrm{t})$, as

$$\bar{\mathbf{y}}(n_\mathrm{t}) = \bar{\mathbf{x}}(n_\mathrm{t}) + \bar{\mathbf{w}}(n_\mathrm{t}) \ , \tag{20}$$

where the noise vector, $\bar{\mathbf{w}}(n_\mathrm{t})$, is defined similar to $\bar{\mathbf{y}}(n_\mathrm{t})$ and

$$\bar{\mathbf{x}}(n_\mathrm{t}) = \begin{bmatrix}\beta_0 s(n_\mathrm{t}-\tau_0) & \cdots & \beta_{N_\mathrm{s}-1} s(n_\mathrm{t}-\tau_{N_\mathrm{s}-1})\end{bmatrix}^T \tag{21}$$

$$= \boldsymbol{\beta}\bar{\mathbf{s}}(n_\mathrm{t}) \ , \tag{22}$$

$$\bar{\mathbf{s}}(n_\mathrm{t}) = \begin{bmatrix}s(n_\mathrm{t}-\tau_0) & \cdots & s(n_\mathrm{t}-\tau_{N_\mathrm{s}-1})\end{bmatrix}^T \ . \tag{23}$$

We derive the CRBs under the assumption that the noise, $\bar{\mathbf{w}}(n_\mathrm{t})$, is complex white Gaussian with zero mean and variance $\sigma^2$. Under this assumption, we can write the log-likelihood function of the observed signal as

$$\begin{aligned}\ln p(\bar{\mathbf{y}};\boldsymbol{\psi}) &= -N\ln(\pi\sigma^2) \\ &\quad - \frac{1}{\sigma^2}\sum_{n_\mathrm{t}=0}^{N_\mathrm{t}-1}\left[\bar{\mathbf{y}}(n_\mathrm{t}) - \boldsymbol{\beta}\bar{\mathbf{s}}(n_\mathrm{t})\right]^H\left[\bar{\mathbf{y}}(n_\mathrm{t}) - \boldsymbol{\beta}\bar{\mathbf{s}}(n_\mathrm{t})\right] \ ,\end{aligned} \tag{24}$$

where

$$\boldsymbol{\psi} = \begin{bmatrix}\omega_0 & \theta & A_1 & \cdots & A_L & \phi_1 & \cdots & \phi_L\end{bmatrix}^T \ . \tag{25}$$

The Fisher information matrix (FIM) for the joint DOA and pitch estimation problem is given by

$$\mathbf{I}(\boldsymbol{\psi}) = -\mathrm{E}\left\{\frac{\partial^2 \ln p(\bar{\mathbf{y}};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right\} \ . \tag{26}$$

If we assume that the covariance matrix of the noise signal does not depend on the parameter vector, $\boldsymbol{\psi}$, the FIM is given by

$$\mathbf{I}(\boldsymbol{\psi}) = \frac{2}{\sigma^2}\mathrm{Re}\left\{\sum_{n_\mathrm{t}=0}^{N_\mathrm{t}-1}\mathbf{D}_{n_\mathrm{t}}^H(\boldsymbol{\psi})\boldsymbol{\beta}^2\mathbf{D}_{n_\mathrm{t}}(\boldsymbol{\psi})\right\} \ , \tag{27}$$

where $\mathrm{Re}\{\cdot\}$ denotes the real part of a complex number, and $\mathbf{D}_{n_\mathrm{t}}(\boldsymbol{\psi})$ is the gradient matrix at time instance $n_\mathrm{t}$ defined as

$$\begin{aligned}\mathbf{D}_{n_\mathrm{t}}(\boldsymbol{\psi}) = \big[&\mathbf{d}_{n_\mathrm{t}}(\omega_0) \quad \mathbf{d}_{n_\mathrm{t}}(\theta) \quad \mathbf{d}_{n_\mathrm{t}}(A_1) \\ &\cdots \quad \mathbf{d}_{n_\mathrm{t}}(A_L) \quad \mathbf{d}_{n_\mathrm{t}}(\phi_1) \quad \cdots \quad \mathbf{d}_{n_\mathrm{t}}(\phi_L)\big] \ .\end{aligned} \tag{28}$$

Note that the columns of $\mathbf{D}_{n_\mathrm{t}}(\boldsymbol{\psi})$ can be interpreted as the gradient vectors with respect to each of the unknown parameters. The gradient vector with respect to the pitch, $\mathbf{d}_{n_\mathrm{t}}(\omega_0)$, is defined as

$$\mathbf{d}_{n_\mathrm{t}}(\omega_0) = \frac{\partial\bar{\mathbf{s}}(n_\mathrm{t})}{\partial\omega_0} \ , \tag{29}$$

and the vectors $\mathbf{d}_{n_\mathrm{t}}(\theta)$, $\mathbf{d}_{n_\mathrm{t}}(A_l)$ and $\mathbf{d}_{n_\mathrm{t}}(\phi_l)$ are defined similar to $\mathbf{d}_{n_\mathrm{t}}(\omega_0)$ for $l = 1, \ldots, L$. The individual entries of the gradient vectors are given by

$$[\mathbf{d}_{n_\mathrm{t}}(\omega_0)]_{n_\mathrm{s}} = \sum_{l=1}^{L} jlA_l \left(n_\mathrm{t} - f_\mathrm{s} n_\mathrm{s} \frac{d\sin\theta}{c}\right) \times e^{jl\omega_0\left(n_\mathrm{t} - f_\mathrm{s} n_\mathrm{s}\frac{d\sin\theta}{c}\right) + j\phi_l} , \quad (30)$$

$$[\mathbf{d}_{n_\mathrm{t}}(\theta)]_{n_\mathrm{s}} = -\sum_{l=1}^{L} jlA_l\omega_0 f_\mathrm{s} n_\mathrm{s} \frac{d\cos\theta}{c} \times e^{jl\omega_0\left(n_\mathrm{t} - f_\mathrm{s} n_\mathrm{s}\frac{d\sin\theta}{c}\right) + j\phi_l} , \quad (31)$$

$$[\mathbf{d}_{n_\mathrm{t}}(A_l)]_{n_\mathrm{s}} = e^{jl\omega_0\left(n_\mathrm{t} - f_\mathrm{s} n_\mathrm{s}\frac{d\sin\theta}{c}\right) + j\phi_l} , \quad (32)$$

$$[\mathbf{d}_{n_\mathrm{t}}(\phi_l)]_{n_\mathrm{s}} = jA_l e^{jl\omega_0\left(n_\mathrm{t} - f_\mathrm{s} n_\mathrm{s}\frac{d\sin\theta}{c}\right) + j\phi_l} , \quad (33)$$

for $n_\mathrm{s} = 0, \ldots, N_\mathrm{s} - 1$. The exact CRB for the $k$th parameter in $\boldsymbol{\psi}$ is defined as the $(k,k)$th element of the inverse FIM, i.e.,

$$\mathrm{CRB}\left([\boldsymbol{\psi}]_k\right) = \left[\mathbf{I}^{-1}(\boldsymbol{\psi})\right]_{kk} . \quad (34)$$

### *B. Asymptotic Bounds*

The exact CRB expressions are rather complicated, and it is difficult to see how the different parameters and the sample lengths influence the different CRBs. Furthermore, it is hard to see from the exact CRB expressions if there are any benefits of estimating the DOA and pitch jointly compared to estimating them separately. Therefore, we also derive simpler asymptotic CRBs for the joint DOA and pitch estimation problem.

The asymptotic bounds are derived under the assumption that the sensors in the ULA are closely spaced such that $\boldsymbol{\beta} \approx \mathbf{I}$. First, we introduce a new variable,

$$\Delta(x,y) = \sum_{n_\mathrm{t}=0}^{N_\mathrm{t}-1} \mathrm{Re}\left\{\mathbf{d}_{n_\mathrm{t}}^H(x)\mathbf{d}_{n_\mathrm{t}}(y)\right\} \quad (35)$$

$$= \Delta(y,x) . \quad (36)$$

For $N_\mathrm{s} \to \infty$ and $N_\mathrm{t} \to \infty$, we know that the frequency spaced sinusoids are orthogonal. For large $N_\mathrm{s}$ and $N_\mathrm{t}$, it follows that

$$\Delta(\omega_0,\omega_0) \approx \left[\frac{N_\mathrm{t}(N_\mathrm{t}-1)(2N_\mathrm{t}-1)}{6}N_\mathrm{s} + N_\mathrm{t}\zeta^2\sin^2\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)(2N_\mathrm{s}-1)}{6} - \frac{N_\mathrm{t}(N_\mathrm{t}-1)}{2}\zeta\sin\theta N_\mathrm{s}(N_\mathrm{s}-1)\right]\sum_{l=1}^{L} l^2A_l^2 , \quad (37)$$

$$\Delta(\omega_0,\theta) \approx \left[-\frac{N_\mathrm{t}(N_\mathrm{t}-1)}{2}\omega_0\zeta\cos\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)}{2} + N_\mathrm{t}\omega_0\zeta^2\frac{\sin 2\theta}{2}\frac{N_\mathrm{s}(N_\mathrm{s}-1)(2N_\mathrm{s}-1)}{6}\right]\sum_{l=1}^{L} l^2A_l^2, \quad (38)$$

$$\Delta(\omega_0,A_l) \approx 0 \quad (39)$$

$$\Delta(\omega_0,\phi_l) \approx \left[\frac{N_\mathrm{t}(N_\mathrm{t}-1)}{2}N_\mathrm{s} - N_\mathrm{t}\zeta\sin\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)}{2}\right]lA_l^2, \quad (40)$$

$$\Delta(\theta,\theta) \approx N_\mathrm{t}\omega_0^2\zeta^2\cos\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)(2N_\mathrm{s}-1)}{6}\sum_{l=1}^{L} l^2A_l^2, \quad (41)$$

$$\Delta(\theta,A_l) \approx 0 , \quad (42)$$

$$\Delta(\theta,\phi_l) \approx -N_\mathrm{t}\omega_0\zeta\cos\theta\frac{N_\mathrm{s}(N_\mathrm{s}-1)}{2}lA_l^2 , \quad (43)$$

$$\Delta(A_p,A_q) = \begin{cases} N_\mathrm{t}N_\mathrm{s}, & p = q \\ (\approx)0, & p \neq q , \end{cases} \quad (44)$$

$$\Delta(A_p,\phi_q) = \begin{cases} 0, & p = q \\ (\approx)0, & p \neq q , \end{cases} \quad (45)$$

$$\Delta(\phi_p,\phi_q) = \begin{cases} N_\mathrm{t}N_\mathrm{s}A_l^2, & p = q \\ (\approx)0, & p = q , \end{cases} \quad (46)$$

where $\zeta = \frac{f_\mathrm{s}d}{c}$. Furthermore, we know that [41]

$$\begin{bmatrix}\mathbf{A} & \mathbf{U} \\ \mathbf{V} & \mathbf{B}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{C}^{-1} & -\mathbf{C}^{-1}\mathbf{U}\mathbf{B}^{-1} \\ -\mathbf{B}^{-1}\mathbf{V}\mathbf{C}^{-1} & \mathbf{B}^{-1}\mathbf{V}\mathbf{C}^{-1}\mathbf{U}\mathbf{B}^{-1} + \mathbf{B}^{-1}\end{bmatrix}, \quad (47)$$

with $\mathbf{C} = \mathbf{A} - \mathbf{U}\mathbf{B}^{-1}\mathbf{V}$. We now apply (47) on the FIM with the expressions in (37)-(46) and with

$$\mathbf{A} = \begin{bmatrix}\Delta(\omega_0,\omega_0) & \Delta(\omega_0,\theta) \\ \Delta(\theta,\omega_0) & \Delta(\theta,\theta)\end{bmatrix} , \quad (48)$$

$$\mathbf{U} = \begin{bmatrix}0 & \Delta(\omega_0,\phi_1) & \cdots & 0 & \Delta(\omega_0,\phi_L) \\ 0 & \Delta(\theta,\phi_1) & \cdots & 0 & \Delta(\theta,\phi_L)\end{bmatrix} \quad (49)$$

$$= \mathbf{V}^H , \quad (50)$$

$$\mathbf{B} = \mathrm{diag}\{[\Delta(A_1,A_1) \quad \Delta(\phi_1,\phi_1) \quad \cdots \quad \Delta(A_L,A_L) \quad \Delta(\phi_L,\phi_L)]\} . \quad (51)$$

The asymptotic CRBs for the DOA and the pitch can then be found from the diagonal elements of the matrix, $\mathbf{C}^{-1}$. Here, we only derive the asymptotic CRBs for these two parameters while the derivations for the other parameters are left to the interested reader. Some tedious manipulations yield

$$\mathrm{CRB}(\omega_0) \approx \frac{6}{N_\mathrm{t}^3N_\mathrm{s}}\mathrm{PSNR}^{-1} , \quad (52)$$

$$\mathrm{CRB}(\theta) \approx \left[\left(\frac{c}{\omega_0 f_\mathrm{s} d\cos\theta}\right)^2\frac{6}{N_\mathrm{t}N_\mathrm{s}^3} + \left(\frac{\tan\theta}{\omega_0}\right)^2\frac{6}{N_\mathrm{t}^3N_\mathrm{s}}\right]\mathrm{PSNR}^{-1} , \quad (53)$$

where

$$\mathrm{PSNR} = \frac{\sum_{l=1}^{L} l^2A_l^2}{\sigma^2} \quad (54)$$

is the so-called pseudo signal-to-noise ratio. In Fig. 2, we see that the asymptotic bounds indeed approaches the exact bounds for large $N_\mathrm{s}$s and $N_\mathrm{t}$s. To obtain the results in Fig. 2, we used the following set up: the pitch was $f_0 = 100$ Hz, the DOA was $\theta = 20°$, the model order was $L = 4$, the variance of the noise was $\sigma^2 = 0.1$, the sampling frequency was $f_\mathrm{s} = 2$ kHz, the wave propagation speed was $c = 340$ m/s, and the inter-element spacing was $d = 2c/f_\mathrm{s}$. Furthermore,

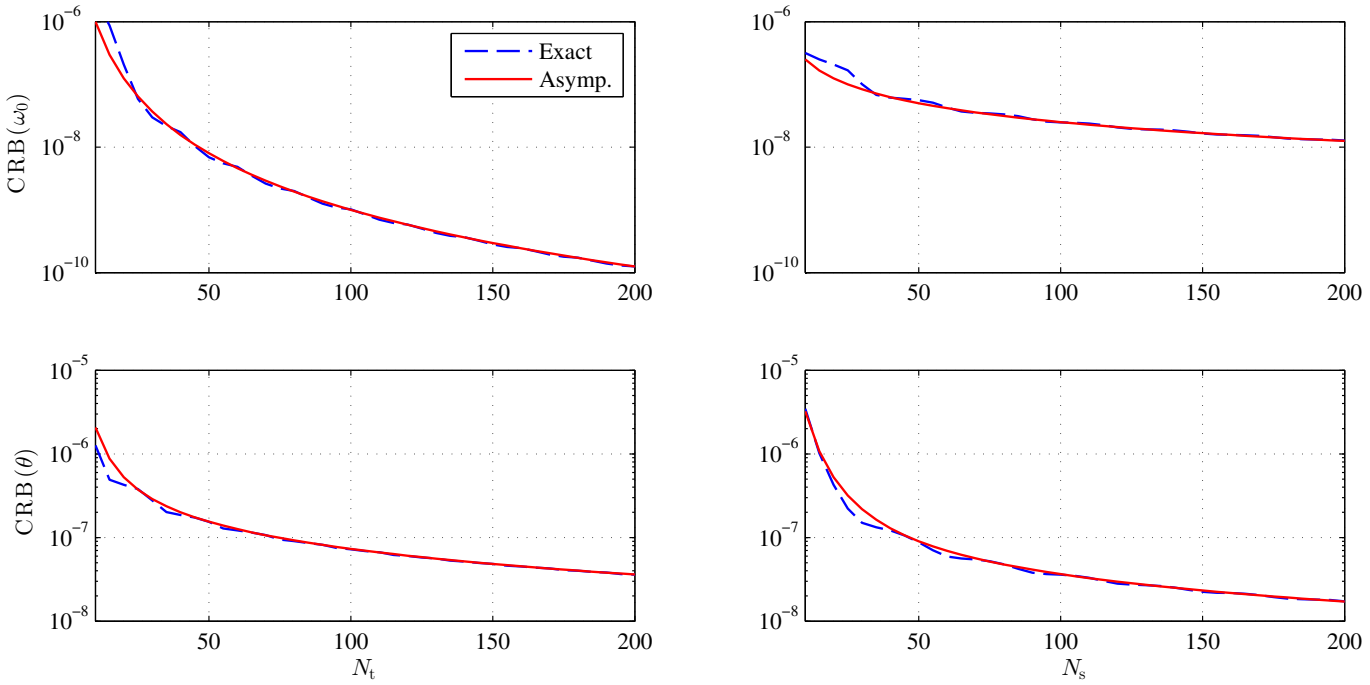


Fig. 2. Plot of the exact and asymptotic Cramér-Rao bounds for (top) the pitch and (bottom) the DOA of the joint DOA and pitch estimation problem as a function of (left) $N_t$ and (right) $N_s$.

the number of sensors was $N_s = 20$ for the simulations with varying $N_t$, and the number of samples was $N_t = 20$ for the simulations with varying $N_s$.

### C. Motivation for Joint DOA and Pitch Estimation

By investigating the asymptotic CRB expressions, it can be seen that the bound for $\omega_0$ is decreasing cubically in $N_t$ and linearly in $N_s$. The bound for $\theta$ consists of two terms; one of the terms is linear in $N_t$ and cubical in $N_s$, and vice versa for the other term. Moreover, it can be seen that it is beneficial to estimate the DOA and the pitch jointly rather then separately. First, we can see from the asymptotic DOA bound in (53) that the CRB is decreased by taking the harmonic signal structure into account as opposed to if we estimated the DOA of a single sinusoid since the bound depends on the PSNR. Moreover, we can see from the asymptotic bound in (52) that the CRB of the pitch can be decreased linearly by increasing the number of sensors, $N_s$.

The DOA and the pitch could also be estimated separately using a two-step procedure where we 1) estimate the DOA and extract the signal impinging from the estimated DOA, and 2) estimate the pitch from the extracted signal. Similarly, we could also estimate the pitch first, extract the signal with the estimated pitch, and then estimate the DOA of the extracted signal. We term such estimation methods as cascaded methods. The cascaded methods, however, will most likely increase the CRBs of the parameters to be estimated in the second step. The cause of the CRB increase is the signal extraction occurring in the first step of the cascaded methods, since the extraction is often performed by a filter which, in general, does not span or contains the subspace spanned by the gradient matrix, $\mathbf{D}_{n_t}(\boldsymbol{\psi})$.

## IV. Joint DOA and Pitch Estimation

In this section, we propose two estimators that jointly estimate the DOA and the pitch of a periodic source that is sampled by a ULA. The methods are based on nonlinear least-squares (NLS), and they are derived under a white Gaussian noise assumption.

### A. Nonlinear Least-Squares Method

First, we derive the NLS method for joint DOA and pitch estimation. The NLS method is derived under the assumption that the noise is white Gaussian. If the noise is indeed white Gaussian, the proposed NLS method resembles the maximum likelihood (ML) estimator, i.e., it will attain the CRB. The proposed NLS method may even provide accurate estimates when the noise is not white Gaussian, as the NLS method for a single sinusoid derived for white Gaussian noise is asymptotically efficient even for colored noise [42].

In this paper, the attenuation matrix $\boldsymbol{\beta}$ is considered as known, i.e., the joint NLS estimates of the DOA and pitch are found by solving

$$\left\{\hat{\theta}, \hat{\omega}_0\right\} = \arg \min_{\boldsymbol{\alpha}, \{\theta,\omega_0\}\in\Theta\times\Omega} \left\|\mathbf{y} - \bar{\mathbf{Z}}\boldsymbol{\alpha}\right\|_2^2 , \qquad (55)$$

with $\|\cdot\|_2$ denoting the $\ell_2$-norm. Minimizing (55) with respect to the complex amplitude vector, $\boldsymbol{\alpha}$, yields

$$\hat{\boldsymbol{\alpha}} = (\bar{\mathbf{Z}}^H\bar{\mathbf{Z}})^{-1}\bar{\mathbf{Z}}^H\mathbf{y} . \qquad (56)$$

If we then insert (56) into (55), we get that

$$\left\{\hat{\theta}, \hat{\omega}_0\right\} = \arg \max_{\{\theta,\omega_0\}\in\Theta\times\Omega} \mathbf{y}^H\bar{\mathbf{Z}}\left(\bar{\mathbf{Z}}^H\bar{\mathbf{Z}}\right)^{-1}\bar{\mathbf{Z}}^H\mathbf{y} , \qquad (57)$$

The above estimator is referred to as the NLS estimator. If we keep only the highest order terms, the complexity of the estimator per point in the search grid $\Theta \times \Omega$ is $\mathcal{O}(NL^2 + L^3)$ where $N = N_\mathrm{t}N_\mathrm{s}$. On basis of (57), we define the NLS cost-functions as

$$\begin{aligned} J_{\text{NLS}}(\theta, \omega_0) &= \left\| \bar{\mathbf{Z}}^H \mathbf{y} \right\|^2_{\left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1}} \\ &= \text{Tr}\left\{ \bar{\mathbf{Z}}^H \mathbf{y}\mathbf{y}^H \bar{\mathbf{Z}} \left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1} \right\} , \end{aligned} \tag{58}$$

with $\| \cdot \|^2_{\mathbf{W}}$ denoting the weighted $\ell_2$-norm and $\mathbf{W}$ is the weighting matrix. Instead of only using a single-data snapshot, $\mathbf{y}$, in the cost-function in (58), we could replace $\mathbf{y}$ by

$$\mathbf{y}_{n_\mathrm{s}}(n_\mathrm{t}) = \text{vec}\left\{ \begin{bmatrix} y_{n_\mathrm{s}}(n_\mathrm{t}) & \cdots & y_{n_\mathrm{s}}(n_\mathrm{t} - M'_\mathrm{t}) \\ \vdots & \ddots & \vdots \\ y_{n_\mathrm{s}+M'_\mathrm{s}}(n_\mathrm{t}) & \cdots & y_{n_\mathrm{s}+M'_\mathrm{s}}(n_\mathrm{t} - M'_\mathrm{t}) \end{bmatrix} \right\} , \tag{59}$$

with $M'_\mathrm{s} = M_\mathrm{s} - 1$, $M'_\mathrm{t} = M_\mathrm{t} - 1$, $M_\mathrm{s} \leq N_\mathrm{s}$, and $M_\mathrm{t} \leq N_\mathrm{t}$. If we then take the expected value, we get

$$\text{E}\left\{ \| \bar{\mathbf{Z}}^H \mathbf{y}_{n_\mathrm{s}}(n_\mathrm{t}) \|^2_{(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}})^{-1}} \right\} = \text{Tr}\left\{ \bar{\mathbf{Z}}^H \mathbf{R} \bar{\mathbf{Z}} \left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1} \right\} . \tag{60}$$

That is, we can also estimate the DOA and pitch jointly by matching the signal model to the covariance matrix, $\mathbf{R}$, of $\mathbf{y}_{n_\mathrm{s}}(n_\mathrm{t})$ as

$$\left\{ \hat{\theta}, \hat{\omega}_0 \right\} = \arg \max_{\{\theta, \omega_0\} \in \Theta \times \Omega} \text{Tr}\left\{ \bar{\mathbf{Z}}^H \mathbf{R} \bar{\mathbf{Z}} \left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1} \right\} . \tag{61}$$

The computational complexity per grid point for the expectation based estimator is $\mathcal{O}(L^2 M + LM^2 + L^3)$ where $M = M_\mathrm{t}M_\mathrm{s}$. Note that even though this complexity looks worse than for the single snapshot NLS estimator it might not be the case in all scenarios since $M \leq N$. In practice, we do not know the exact covariance matrix $\mathbf{R}$, but we can replace it by, e.g., the sample covariance matrix estimate defined as [19]

$$\hat{\mathbf{R}} = \sum_{m_\mathrm{s}=0}^{N_\mathrm{s}-M_\mathrm{s}} \sum_{m_\mathrm{t}=0}^{N_\mathrm{t}-M_\mathrm{t}} \frac{\mathbf{y}_{m_\mathrm{s}}(n_\mathrm{t} - m_\mathrm{t}) \mathbf{y}^H_{m_\mathrm{s}}(n_\mathrm{t} - m_\mathrm{t})}{(N_\mathrm{s} - M'_\mathrm{s})(N_\mathrm{t} - M'_\mathrm{t})} . \tag{62}$$

The gradient of the cost-function, $J_{\text{NLS}}(\theta, \omega_0)$, is given by

$$\nabla J_{\text{NLS}}(\theta, \omega_0) = \begin{bmatrix} \frac{\partial J_{\text{NLS}}}{\partial \theta} & \frac{\partial J_{\text{NLS}}}{\partial \omega_0} \end{bmatrix}^T , \tag{63}$$

where

$$\frac{\partial J_{\text{NLS}}}{\partial \theta} = \mathbf{y}^H \left( \mathbf{G}_\theta \mathbf{P}^\perp + \mathbf{P}^\perp \mathbf{G}_\theta^H \right) \mathbf{y} , \tag{64}$$

$$\frac{\partial J_{\text{NLS}}}{\partial \omega_0} = \mathbf{y}^H \left( \mathbf{G}_{\omega_0} \mathbf{P}^\perp + \mathbf{P}^\perp \mathbf{G}_{\omega_0}^H \right) \mathbf{y}, \tag{65}$$

with

$$\mathbf{P}^\perp = (\mathbf{I} - \mathbf{P}), \tag{66}$$

$$\mathbf{P} = \bar{\mathbf{Z}} (\bar{\mathbf{Z}}^H \bar{\mathbf{Z}})^{-1} \bar{\mathbf{Z}}^H , \tag{67}$$

$$\mathbf{G}_\theta = \bar{\mathbf{Z}} \left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1} \mathbf{Y}_\theta^H \mathbf{B} , \tag{68}$$

$$\mathbf{G}_{\omega_0} = \bar{\mathbf{Z}} \left(\bar{\mathbf{Z}}^H \bar{\mathbf{Z}}\right)^{-1} \mathbf{Y}_{\omega_0}^H \mathbf{B} , \tag{69}$$

$$[\mathbf{Y}_\theta]_{pq} = -jq\omega_0 \zeta \cos\theta k_{\mathrm{s},p} e^{-jq\omega_0(\zeta \sin\theta k_{\mathrm{s},p} + k_{\mathrm{t},p})} , \tag{70}$$

$$[\mathbf{Y}_{\omega_0}]_{pq} = -jq \left(\zeta \sin\theta k_{\mathrm{s},p} + k_{\mathrm{t},p}\right) e^{-jq\omega_0(\zeta \sin\theta k_{\mathrm{s},p} + k_{\mathrm{t},p})} , \tag{71}$$

$k_{\mathrm{s},p} = (p-1) \pmod{M_\mathrm{s}}$ and $k_{\mathrm{t},p} = \lfloor \frac{p-1}{M_\mathrm{s}} \rfloor$. Note that $y \pmod{x}$ denotes that $y$ is modulo $x$, and $\lfloor \cdot \rfloor$ is the floor operator. Using the gradient in (63), we can iteratively obtain refined estimates of the DOA and the pitch as

$$\begin{bmatrix} \hat{\theta}^{(i+1)} \\ \hat{\omega}_0^{(i+1)} \end{bmatrix} = \begin{bmatrix} \hat{\theta}^{(i)} \\ \hat{\omega}_0^{(i)} \end{bmatrix} + \delta \nabla J_{\text{NLS}} , \tag{72}$$

where $i$ is the iteration index and $\delta > 0$ is a small constant which can be found using a line search algorithm.

### B. Approximate Nonlinear Least-Squares Method

When the number of spatial and temporal samples are large, the harmonics are close to being orthogonal, i.e., [9]

$$\lim_{M \to \infty} \frac{1}{M} \mathbf{Z}^H \mathbf{Z} = \mathbf{I} . \tag{73}$$

Therefore, cf. (14), we know that

$$\lim_{M \to \infty} \frac{1}{M} \bar{\mathbf{Z}}^H \bar{\mathbf{Z}} = \frac{\|\boldsymbol{\beta}\|_2^2}{M_\mathrm{s}} \mathbf{I} . \tag{74}$$

Inserting (74) into (57) yields the approximate NLS (aNLS) estimator defined as

$$\left\{ \hat{\theta}, \hat{\omega}_0 \right\} = \arg \max_{\{\theta, \omega_0\} \in \Theta \times \Omega} \mathbf{y}^H \bar{\mathbf{Z}} \bar{\mathbf{Z}}^H \mathbf{y} . \tag{75}$$

That is, the aNLS cost-function is given by

$$J_{\text{aNLS}}(\theta, \omega_0) = \| \bar{\mathbf{Z}}^H \mathbf{y} \|_2^2 . \tag{76}$$

The computational complexity per search grid point of the aNLS estimator is $\mathcal{O}(NL)$, i.e., it is only quadratic compared to the complexity of the NLS estimator which was cubic[1]. As for the NLS method, we also propose an alternative covariance-based estimator

$$\left\{ \hat{\theta}, \hat{\omega}_0 \right\} = \arg \max_{\{\theta, \omega_0\} \in \Theta \times \Omega} \text{Tr}\left\{ \bar{\mathbf{Z}}^H \mathbf{R} \bar{\mathbf{Z}} \right\} \tag{77}$$

The computational complexity of evaluating the expectation based aNLS estimator in each point of the search grid is $\mathcal{O}(M^2 L + ML^2)$. Note that the alternative aNLS cost-function in (77) can be interpreted as the output power of a periodogram-based filterbank when $\mathbf{B} = \mathbf{I}$.

The expressions for the partial derivatives of the aNLS cost-function are given by

$$\frac{\partial J_{\text{aNLS}}}{\partial \theta} = \mathbf{y}^H \left( \mathbf{B} \mathbf{Y}_\theta \bar{\mathbf{Z}}^H + \bar{\mathbf{Z}} \mathbf{Y}_\theta^H \mathbf{B} \right) \mathbf{y} , \tag{78}$$

$$\frac{\partial J_{\text{aNLS}}}{\partial \omega_0} = \mathbf{y}^H \left( \mathbf{B} \mathbf{Y}_{\omega_0} \bar{\mathbf{Z}}^H + \bar{\mathbf{Z}} \mathbf{Y}_{\omega_0}^H \mathbf{B} \right) \mathbf{y} . \tag{79}$$

We can then obtain refined aNLS estimates by using (78) and (79) in (72).

## V. Experimental Results

To evaluate the proposed joint DOA and pitch estimators, we conducted simulations on both synthetic as well as real-life data. The results from these simulations are explained in the following subsections.

[1]Here, we consider all unknown variables as one variable when counting the order, i.e., $\mathcal{O}(NL)$ is considered as a second order term.

### A. Statistical Evaluation

We conducted several series of Monte-Carlo simulations using synthetic data. In all of these simulations, the sampling frequency was $f_{\rm s} = 8$ kHz, the speed of sound was assumed to be $c = 343.2$ m/s, the array was uniform and linear with $d = \frac{c}{f_{\rm s}}$, there was no attenuation across the sensors such that $\mathbf{B} = \mathbf{I}$, and the desired signal was designed to be a harmonic signal with $f_0 = 243$ Hz, $\theta = 15^\circ$, $L = 5$ and $\alpha_l = 1$. We estimated the pitch and DOA in each of the simulations using different estimators including the proposed; besides the proposed estimators, we used the multichannel maximum likelihood (MC-ML) and multichannel approximate maximum likelihood (MC-aML) estimators [43] for pitch estimation, and we used the steered response power (SRP) method, the steered response power with phase transform (SRP-PHAT) method [44], [45], and the broadband MVDR (bMVDR) beamformer [17] for DOA estimation. Finally, we used the position-pitch plane (PoPi) based estimator in [32], the subspace (Sub.) method in [30], and the LCMV filtering (LCMV) method in [35] for joint DOA and pitch estimation. For pitch estimation, we compared with the MC-ML and MC-aML estimators since these were shown to outperform the multi-channel pitch estimators in [46], [47]. Note that, in our implementations of the SRP and bMVDR methods, we used an FFT length of 256, and we integrated over all frequency indices whereas in the SRP-PHAT method we integrated over the frequency indices corresponding to the interval $[200 \text{ Hz}; L\max\{f_{0,\text{grid}}\}]$ with $\max\{f_{0,\text{grid}}\}$ being our maximum pitch candidate. Moreover, in our implementation of the bMVDR method, we used 20 blocks of length $\lfloor N_{\rm t}/3 \rfloor$ to estimate the cross-spectral density. We used an FFT size of 1024 for the PoPi method, a block size of $N_{\rm t}/2$ and a smoothing factor of 5 for the subspace method, and spatial and temporal filter lengths of $\max\{[2, \lfloor N_{\rm s} \cdot 2/3 \rfloor]\}$ and $\lfloor N_{\rm t}/4 \rfloor$, respectively, for the LCMV method.

In each series of Monte-Carlo simulations, the performances of the estimators were measured in terms mean squared error (MSE). In the first series of Monte-Carlo simulations, we measured the estimation performance as a function of the signal-to-noise ratio (SNR) defined as

$$\text{SNR} = 10\log_{10}\frac{\sum_{l=1}^{L}|\alpha_l|^2}{\sigma^2} \ . \tag{80}$$

The number of sensors were $N_{\rm s} = 2$ and the number of temporal samples was $N_{\rm t} = 80$. Then, we conducted another series of Monte-Carlo simulations where the estimation performance was evaluated versus the number of sensors $N_{\rm s}$ while the SNR was fixed to 10 dB and the number of temporal samples was $N_{\rm t} = 60$. In the third series, we measured the performance as a function of the number of temporal samples $N_{\rm t}$, and, here, the SNR was 30 dB while the number of sensors was $N_{\rm s} = 2$. Finally, we conducted two series of Monte-Carlo simulations on synthetic data containing two harmonic sources each with five harmonics with unit amplitudes. In the first of these series, both sources had a DOA of $15^\circ$. One of the sources had a pitch of 243 Hz, while the pitch of the other source was varied. The MSEs was then measured as the mean of the MSEs for the two sources. For this experiment, the number of sensors and samples was $N_{\rm s} = 2$ and $N_{\rm t} = 80$, respectively. In the other series, the pitch of both sources was 243 Hz, and the DOA of one of the sources was $15^\circ$, while the DOA of the other source was varying. The number of sensors and samples was $N_{\rm s} = 8$ and $N_{\rm t} = 60$, respectively.

The results from all of the series of Monte-Carlo simulations are depicted in Fig. 3, and they reveal several interesting facts. First, we note that the proposed NLS estimator attains the CRB for both the DOA and pitch when the noise is white Gaussian and we have a single harmonic source. This was also expected according to our previous claims. Moreover, the NLS estimator has a better or similar performance than all other methods in the comparison. The proposed aNLS estimator, however, is slightly biased and does therefore not attain the CRB, but in many scenarios it closely follows it. The PoPi, MC-aML, SRP and SPR-PHAT methods are also biased; therefore, as for the aNLS method, the MSEs of their estimates do not necessarily follow the CRB for increasing SNRs, $N_{\rm t}$s or $N_{\rm s}$s. Another key observation for the single-source experiments is that the aNLS method seems to outperform the MC-aML method in most scenarios in terms of the MSE of the pitch estimates.

In the first two-source scenario[2], the DOAs of the sources were the same while the pitch spacing was varying. The NLS, aNLS, MC-ML and MC-aML methods outperform the PoPi and LCMV methods for pitch estimation for pitch spacings above $\approx 0.0155$ in this scenario. Moreover, the proposed NLS and aNLS estimators clearly outperforms all other methods for DOA estimation. It is expected that the SRP, SRP-PHAT, and bMVDR methods fail in this scenario, as the broadband methods can not resolve sources with the same DOA. In the other two-source scenario, the two sources had the same pitch, while the DOA spacing between the sources was altered. Here, we observe that the proposed methods outperform all other methods for pitch estimation for DOA spacings below $\approx -0.87$. We note that the MC-ML and MC-aML methods fail in this scenario, since they only conduct a one-dimensional search. For DOA estimation, the NLS, aNLS and SRP methods yields the best performance for DOA spacings below $\approx -0.87$.

### B. Real-life Examples

We also conducted some qualitative experiments to evaluate the performance of the proposed methods on real-life signals. These experiments were conducted in a meeting room. The floor plan of the room and the measurement setup are illustrated in Fig. 4, while the height of the room was 2.64 m. In these simulations, the sampling frequency was $f_{\rm s} = 44.1$ kHz, the speed of sound was assumed to be $c = 343.2$ m/s, the room reverberation time[3] at 1 kHz was $T_{60} \approx 0.53$ s, the array was uniform and linear with $d = 4$ cm and $N_{\rm s} = 8$, we assume that there was no attenuation across the sensors such that $\mathbf{B} = \mathbf{I}$, and the desired signal was assumed to consist of $L = 8$ harmonics. The estimators used in these experiments were the same as in the previous simulations with synthetic data and they were set up similarly.

[2]The subspace method is not considered in these scenarios as it is only suited for estimating the parameters of a single source.

[3]Here, the reverberation time is defined as time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound.

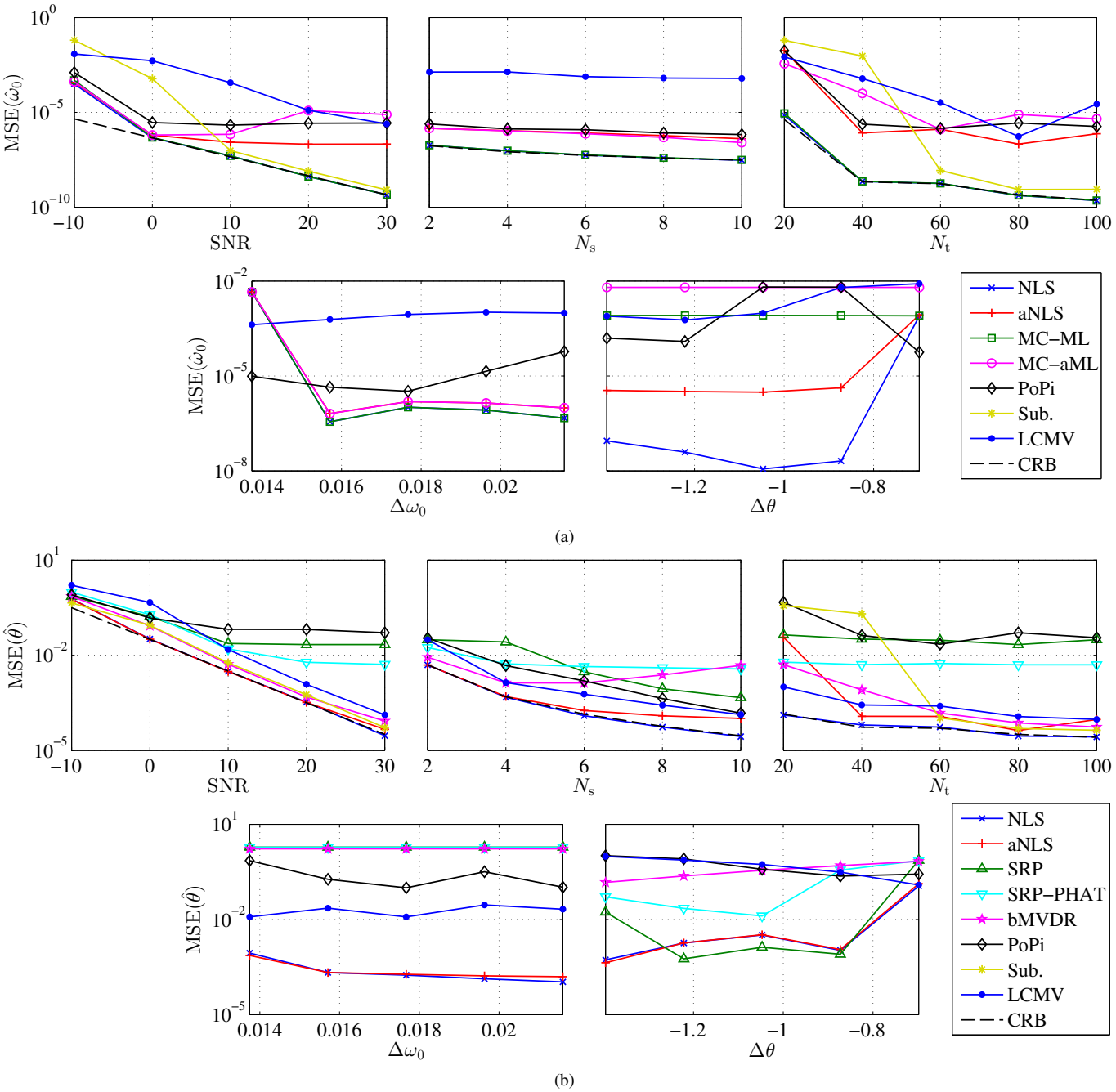


Fig. 3. MSE of (a) pitch and (b) DOA estimates obtained in different scenarios with the NLS and aNLS methods proposed herein, the multichannel maximum likelihood method (MC-ML), the approximate MC-ML method (MC-aML), the steered response power method (SRP), the SRP with phase transform method (SRP-PHAT), the broadband MVDR beamformer (bMVDR), the position-pitch plane based method (PoPi), a subspace based method (Sub.), and an LCMV filtering method (LCMV). In all scenarios, 500 Monte-Carlo simulations were conducted for each experimental setup to estimate the MSE of the parameter estimates.

In the first of the real-life experiments, we played back an anechoic trumpet signal using the speaker 'S2'. The anechoic trumpet signal was generated by concatenating anechoic trumpet signal excerpts[4]. The played back trumpet signal was recorded using the ULA to obtain a multichannel trumpet signal with slight reverberation. From the recording, we estimated the pitch and the DOA of the trumpet signal using the estimators mentioned previously. In Fig. 5, the estimates obtained from this experiment are depicted. We can see that all of the applied estimators for pitch estimation except the PoPi and LCMV methods seem to correctly estimate the pitch of the trumpet signal if we compare the estimates with the spectrogram. Regarding DOA estimation, we can see that all estimators obtain estimates relatively close to the true DOA except the bMVDR and LCMV methods which looks heavily biased. Note that the NLS and aNLS methods yield estimates close to the true parameter values even though the recording contains reverberation and $\mathbf{B} \neq \mathbf{I}$ in practice. We

[4]The excerpts were downloaded from http://theremin.music.uiowa.edu/MIS.html

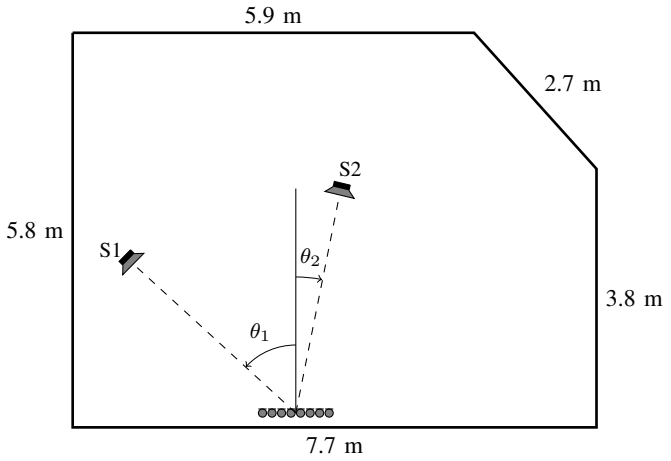


Fig. 4. Floor plan of the meeting room used for the real-life experiments. The angle between the two speakers, 'S1' and 'S2', was $\theta_1 \approx 45°$ and $\theta_2 \approx -13°$, respectively.

then conducted a similar experiment where we played back the same trumpet signal using 'S1' and a speech signal using 'S2', and the mixture was recorded using the array. The played back speech signal was a female speech excerpt taken from the Keele pitch database [48]. In this experiment, the speech signal was considered noise, so the objective was to estimate the DOA and the pitch of the trumpet signal. In Fig. 6, the results from this experiment are shown. Again, we observe that the pitch estimators except for the PoPi and LCMV estimators seem to provide correct pitch estimates except at $\approx$ 7.3 s. For DOA estimation, it seems that the proposed methods outperform the other methods. The SRP-PHAT method provides heavily biased estimates, and the estimates of the bMVDR, PoPi, and LCMV methods seem erroneous. We note that the proposed methods provide good pitch and DOA estimates even though the noise is indeed not white Gaussian in this experiment. In summary, the proposed methods show comparable or better estimation performance than other state-of-the-art DOA and pitch estimators in our real-life experiments. Moreover, the results from these experiments indicate that the proposed methods are applicable on real-life signals, and that they are robust against reverberation as well as other noise types than white Gaussian noise. Note that the above observations based on our qualitative experiments may be different for, e.g., other sensor and source positions, and array structures, due to the complicated nature of reverberant signals.

## VI. Conclusion

In this paper, we have considered joint estimation of the DOA and the pitch of a harmonic source recorded using a ULA. First, we derived the exact and asymptotic Cramér-Rao bounds (CRBs) for the joint estimation problem. From the asymptotic bounds, it is clear that the DOA can be estimated more accurately by taking the harmonic structure into account compared to if we just estimated the DOA of, e.g., the fundamental tone. Moreover, these bounds reveal that the pitch can be estimated more accurately when multiple sensors are used. Then, we proposed two estimators for joint DOA and pitch estimation, namely the NLS and aNLS methods. The proposed estimators are maximum likelihood estimators when the noise is white Gaussian, the environment is anechoic, and the source of interest is in the far-field. We conducted numerous simulations on synthetic data where

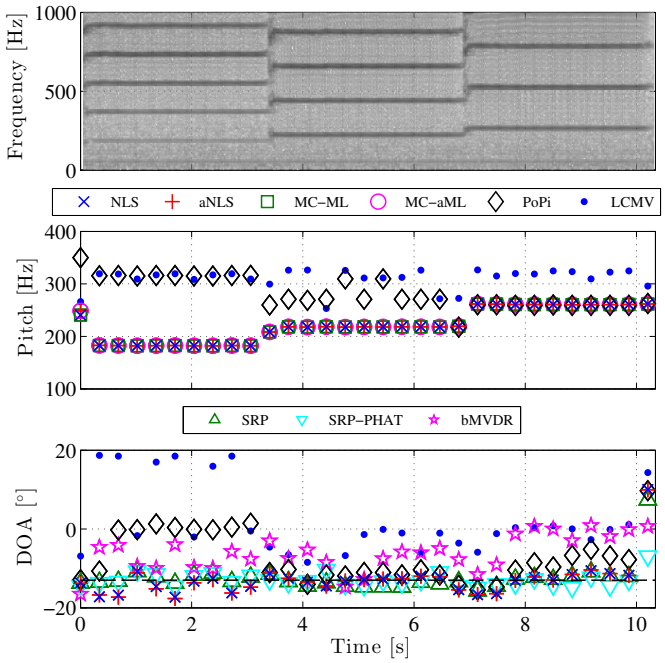


Fig. 5. Estimation results from a real-life experiment with a single source; the source was a trumpet signal played back using 'S2'. The pitch and DOA estimates of the trumpet signal is depicted in the top and bottom plots, respectively.

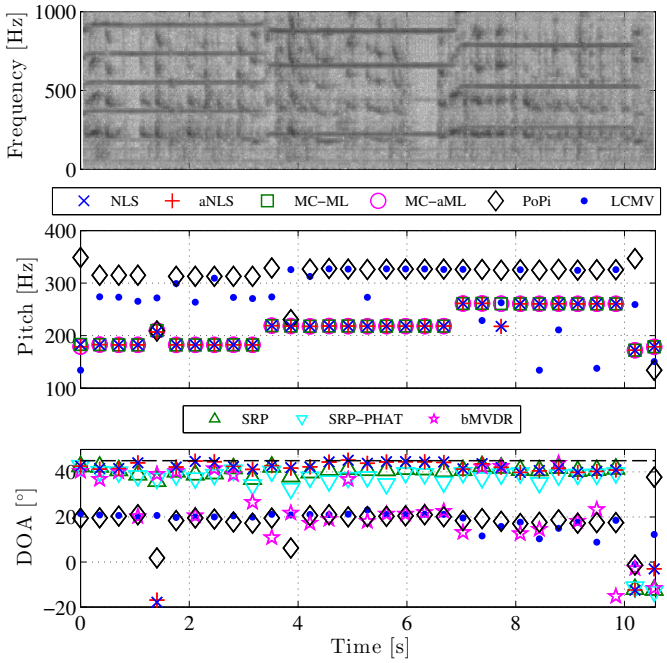


Fig. 6. Estimation results from a real-life experiment with two sources; the sources were a trumpet signal and a speech signal played back using 'S1' and 'S2', respectively. The pitch and DOA estimates of the trumpet signal is depicted in the top and bottom plots, respectively.

the proposed methods and other state-of-the-art methods for DOA and pitch estimation were applied. The results show that the proposed methods attain the CRB with the aNLS being slightly biased. In general, the proposed methods outperform the other methods for both DOA and pitch estimation in terms of mean squared error. This is even the case in two-source scenarios, where the noise is not white Gaussian. The results obtained from the two-source scenarios also show that it is beneficial to estimate the DOAs and the pitches jointly when two sources are having the same DOA or pitch, since the methods estimating only one of these parameters may fail. Furthermore, we conducted experiments on real-life data. The results from these experiments indicate that the proposed methods have similar or better estimation performance than the other applied methods. Moreover, these experiments indicate that the proposed methods are applicable on real-life signals, and that they are robust against reverberation and noise which is not white Gaussian. In future work, model order estimation for the methods proposed herein can be considered, to improve their robustness and applicability.

## 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] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, Inc., 2002.

[6] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[7] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis - Principles, Algorithm, and Applications*. John Wiley & Sons, Inc., 2006.

[8] S. Makino, T. W. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

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

[10] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. Signal Process.*, vol. 39, no. 1, pp. 40–48, Jan. 1991.

[11] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 5, pp. 1635–1644, Jul. 2007.

[12] V. Emiya, B. David, and R. Badeau, "A parametric method for pitch estimation of piano tones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2007, pp. 249–252.

[13] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Elsevier Signal Process.*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[14] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, Jul. 2003.

[15] L. Du, T. Yardibi, J. Li, and P. Stoica, "Review of user parameter-free robust adaptive beamforming algorithms," *Digital Signal Processing*, vol. 19, no. 4, pp. 567–582, Jul. 2009.

[16] M. Viberg and P. Stoica, "A computationally efficient method for joint direction finding and frequency estimation in colored noise," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 2, Nov. 1998, pp. 1547–1551.

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

[18] ——, *Nonlinear Methods of Spectral Analysis*. Springer-Verlag, 1983, ch. Maximum-Likelihood Spectral Estimation.

[19] A. Jakobsson, S. L. Jr. Marple, and P. Stoica, "Computationally efficient two-dimensional Capon spectrum analysis," *IEEE Trans. Signal Process.*, vol. 48, no. 9, pp. 2651–2661, Sep. 2000.

[20] A. N. Lemma, A.-J. van der Veen, and E. F. Deprettere, "Analysis of joint angle-frequency estimation using ESPRIT," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1264–1283, May 2003.

[21] T. Shu and X. Liu, "Robust and computationally efficient signal-dependent method for joint DOA and frequency estimation," *EURASIP J. on Advances in Signal Processing*, vol. 2008, no. 1, pp. 1–16, Apr. 2008.

[22] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Elsevier Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[23] 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.

[24] 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.

[25] G. C. Carter, "Coherence and time delay estimation," *Proc. IEEE*, vol. 75, no. 2, pp. 236–255, Feb. 1987.

[26] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 1997, pp. 375–378.

[27] M. Jian, A. C. Kot, and M. H. Er, "DOA estimation of speech source with microphone arrays," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 5, May 1998, pp. 293–296.

[28] X. Qian and R. Kumaresan, "Joint estimation of time delay and pitch of voiced speech signals," *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 1, pp. 735–739, Oct. 1995.

[29] G. Liao, H. C. So, and P. C. Ching, "Joint time delay and frequency estimation of multiple sinusoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May 2001, pp. 3121–3124.

[30] L. Y. Ngan, Y. Wu, H. C. So, P. C. Ching, and S. W. Lee, "Joint time delay and pitch estimation for speaker localization," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 3, May 2003, pp. 722–725.

[31] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–11, Jan. 2012.

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

[33] M. Képesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, May 2008, pp. 85–88.

[34] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 8, pp. 1490–1502, Nov. 2008.

[35] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint DOA and fundamental frequency estimation methods based on 2-d filtering," in *Proc. European Signal Processing Conf.*, Aug. 2010, pp. 2091–2095.

[36] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[37] S. Godsill and M. Davy, "Bayesian computational models for inharmonicity in musical instruments," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2005, pp. 283–286.

[38] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, "Robust subspace-based fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 101–104.

[39] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[40] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Inc., 1993.

[41] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. The John Hopkins University Press, 1996.

[42] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case: asymptotic Cramér-Rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Trans. Signal Process.*, vol. 45, no. 8, pp. 2048–2059, Aug. 1997.

[43] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.
[44] 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. Springer-Verlag, 2001, ch. 8, pp. 157–180.
[45] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, May 2000.
[46] L. Armani and M. Omologo, "Weighted autocorrelation-based f0 estimation for distant-talking interaction with a distributed microphone network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2004, pp. 113–116.
[47] F. Flego and M. Omologo, "Robust f0 estimation based on a multi-microphone periodicity function for distant-talking speech," in *Proc. European Signal Processing Conf.*, Sep. 2006, pp. 1–4.
[48] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

**Jesper Rindom Jensen** (S'09–M'12) was born in Ringkøbing, Denmark in August 1984. He received the M.Sc. degree cum laude for completing the elite candidate education in 2009 from Aalborg University in Denmark. In 2012, he received the Ph.D. degree from Aalborg University.

Currently, he is a Postdoctoral Researcher at the Department of Architecture, Design & Media Technology at Aalborg University in Denmark, where he is also a member of the Audio Analysis Lab. He has been a Visiting Researcher at University of Quebec, INRS-EMT, in Montreal, Quebec, Canada. He has published several papers in peer-reviewed conference proceedings and journals. Among others, his research interests are digital signal processing and microphone array signal processing theory and methods with application to speech and audio signals. In particular, he is interested in parametric analysis, modeling and extraction of such signals.

**Mads Græsbøll Christensen** (S'00–M'05–SM'11) was born in Copenhagen, Denmark, in March 1977. He received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University (AAU) in Denmark, where he is also currently employed at the Dept. of Architecture, Design & Media Technology as Associate Professor. At AAU, he is head of the Audio Analysis Lab which conducts research in audio signal processing.

He was formerly with the Dept. of Electronic Systems, Aalborg University and has been a Visiting Researcher at Philips Research Labs, ENST, UCSB, and Columbia University. He has published more than 100 papers in peer-reviewed conference proceedings and journals as well as 1 research monograph. His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Dr. Christensen has received several prestigious awards and grants, including an ICASSP Student Paper Award, the Spar Nord Foundation's Research Prize, a Danish Independent Research Council Young Researcher's Award and postdoc grant, and a Villum Foundation Young Investigator Programme grant. He is an Associate Editor for the IEEE Signal Processing Letters.

**Søren Holdt Jensen** (S87–M88–SM00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. Before joining the Department of Electronic Systems of Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd, Copenhagen, Denmark; the Electronics Institute of the Technical University of Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education (UNI•C), Lyngby; the Electrical Engineering Department of Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK) of Aalborg University. He is Full Professor and is currently heading a research team working in the area of numerical algorithms, optimization, and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications.

Prof. Jensen was an Associate Editor for the IEEE Transactions on Signal Processing and Elsevier Signal Processing, and is currently Associate Editor for the IEEE Transactions on Audio, Speech and Language Processing. He is a recipient of an European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section and the IEEE Denmark Sections Signal Processing Chapter. He is member of the Danish Academy of Technical Sciences and was in January 2011 appointed as member of the Danish Council for Independent Research–Technology and Production Sciences by the Danish Minister for Science, Technology and Innovation.