Aalborg Universitet

AALBORG UNIVERSITY

Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation

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

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

DOI (link to publication from Publisher):
10.1109/TASLP.2014.2377583

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Jensen, J. R., Christensen, M. G., Benesty, J., & Jensen, S. H. (2015). Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation. *I E E E Transactions on Audio, Speech and Language Processing*, *23*(1), 174-185. https://doi.org/10.1109/TASLP.2014.2377583

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: June 18, 2025

Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation

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

***Abstract*—In this paper, spatio-temporal filtering methods are proposed for estimating the direction-of-arrival (DOA) and fundamental frequency of periodic signals, like those produced by the speech production system and many musical instruments using microphone arrays. This topic has quite recently received some attention in the community and is quite promising for several applications. The proposed methods are based on optimal, adaptive filters that leave the desired signal, having a certain DOA and fundamental frequency, undistorted and suppress everything else. The filtering methods simultaneously operate in space and time, whereby it is possible resolve cases that are otherwise problematic for pitch estimators or DOA estimators based on beamforming. Several special cases and improvements are considered, including a method for estimating the covariance matrix based on the recently proposed iterative adaptive approach (IAA). Experiments demonstrate the improved performance of the proposed methods under adverse conditions compared to the state of the art using both synthetic signals and real signals, as well as illustrate the properties of the methods and the filters.**

***Index Terms*—2-D filtering, DOA estimation, fundamental frequency estimation, joint estimation, LCMV beamformer, periodogram-based beamformer.**

I. Introduction

A FUNDAMENTAL property of speech and audio signals is the so-called pitch. For many signals, namely periodic signals, the pitch is equivalent to the fundamental frequency, i.e., the frequency of which integer multiples form the frequencies of the individual harmonics, even though there exists some pathological examples where it is not quite that simple. In some applications, the pitch itself is of interest or is being studied for other purposes, some examples being prosody analysis and transcription of music. The pitch also often forms the basis of the processing of such signals. Some well-known examples include speech coding, wherein long-term predictors are used to exploit the correlation caused by the quasi-periodicity that causes the pitch, and noise reduction, wherein the pitch can be used to either directly enhance the signal of interest [1] or to estimate the properties of the noise [2]. Filters that extract or attenuate the harmonics of periodic sounds are often referred to as comb filters, due to their characteristic frequency response. Such comb filters have played a prominent role in the history of signal processing, dating back to 1970's [3], and new forms of comb filters keep emerging. The classical comb filter is based on signal-independent FIR or IIR filters with poles or zeros, respectively, close to the unit circle at the harmonic frequencies. Later, it was shown that more efficient filters can be obtained via a set of notch or peak filters [4], and a few other examples of such approaches can be found in [5] and the references therein. More recently, it was shown that by generalizing the principle of the Capon spectral estimator, it is possible to design optimal, adaptive FIR comb filters [1], [6]. These filters have a number of properties that make them desirable in several applications. The filters are distortion-less, i.e., they let the signal of interest, i.e., periodic signals, pass undistorted. They are adaptive and, hence, automatically adapt to the conditions under which the signal of interest has been recorded. This means that they can cancel strong interferences, including also other periodic signals, without prior knowledge of their properties. The filters also, curiously, reduce to evaluating Fourier transforms at certain frequencies or projecting onto the space spanned by Fourier bases under certain conditions.

In microphone arrays, the direction-of-arrival (DOA) is often used as a means of locating, tracking and separating signals, something that is often done using spatial filters, i.e., beamformers [7], [8]. Since speech and audio signals are generally broadband, unlike, for example, communication and radar signals, many of the clever narrowband beamforming techniques cannot be applied directly to such signals. Instead, speech and audio signals are often decomposed into a set of subbands, each of which are then processed as narrowband signals. However, periodic signals can be modeled efficiently using the harmonic model [9], in which the signal of interest is modeled as a set of narrowband signals, namely sinusoids corresponding to the individual harmonics. This means that such signals can in fact be treated as multiple narrowband signals that share some common parameters: the fundamental frequency and the DOA. In fact, by finding, jointly, both the fundamental frequency (i.e., the pitch) and the DOA, it is possible to mitigate some of the severe problems that pitch estimators encounter for multiple sources, and it is possible to overcome some of the problems that DOA estimators have with distinguishing between different sources when these impinge on the array from angles that

Manuscript received February 24, 2014; revised June 27, 2014; accepted November 25, 2014. Date of publication December 08, 2014; date of current version January 14, 2015. This work was supported by the Villum Foundation and the Danish Council for Independent Research under Grant DFF-1337-00084. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nobutaka Ono.

J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, AD:MT, Aalborg University, DK-9000 Aalborg, Denmark (e-mail: jrj@create.aau.dk; mgc@create.aau.dk).

J. Benesty is with the INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada, and also with the Audio Analysis Lab, AD:MT, Aalborg University, Denmark (e-mail: benesty@emt.inrs.ca).

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2377583

2329-9290 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

are close. It should also be noted that the DOA along with the pitch also are believed to be some of the governing factors that the human auditory system uses for separating sources. This line of reasoning has, quite recently, led to some joint DOA and fundamental frequency estimators, including maximum likelihood based [10], [11], subspace-based [12]–[14], correlation-based [15], [16], and filtering-based [17]–[19] methods. Notably, the problem of joint DOA and fundamental frequency estimation was formalized and thoroughly analyzed in [10], and a maximum likelihood estimator that achieves the highest possible accuracy (under certain conditions) was proposed.

In this paper, we propose spatio-temporal filtering methods for joint DOA and fundamental frequency estimation for periodic signals, like, for example, speech signals or signals produced by musical instruments. The filters are based on the principle of the Capon and Frost beamformers [20], [21] and spectral estimators combined and generalized to account for the nature of periodic signals, and they are controlled by two parameters: the DOA and the fundamental frequency. The proposed filters are optimal and adaptive, and should, hence, be capable of dealing with adverse conditions, like when strong background noise or interference is present but guarantee that the signal of interest is left undistorted. The filters can be thought of as jointly performing beamforming and enhancement, i.e., they are spatio-temporal. In this paper, however, we consider only the application of these filters to parameter estimation, i.e., estimation of the DOA and the fundamental frequency. We consider various variations and simplifications of the filters, including optimal filters for white noise and for infinitely long filters. We also consider the application of the principle of the iterative adaptive approach (IAA) [19], [22], [23] for finding the covariance matrix, which is required to compute the optimal filters. This can be used to obtain longer filters for a given number of samples, something that often results in an improved estimation of parameters, especially under adverse conditions. While the IAA based estimators are computationally more complex than using the traditional sample covariance matrix estimate, it has been shown [24], [25] that the computational complexity of the IAA can be reduced dramatically.

The rest of the paper is organized as follows: In Section II, we introduce the problem formulation along with some useful notation, and we proceed to motivate the usage of joint DOA and fundamental frequency estimation in more detail. We then, in Section III, introduce the filter designs and consider, as mentioned, various special cases and the IAA method for estimation of the covariance matrix. In Section IV, the experimental results are presented, after which the conclusion follows in Section V.

II. Problem Statement

Consider a scenario where K microphones are recording a mixture of a desired, noise, and interfering sources. At time instance n, we can then model the signal observed using the k'th microphone as

$$y_k(n) = x_k(n) + v_k(n), \tag{1}$$

where $x_k(n)$ is the recording of the desired source, and $v_k(n)$ is the sum of the recorded noise and interference. In this paper, we assume that the desired signal is periodic, which is a reasonable assumption for, e.g., voiced speech and many musical instruments. The noise can, for example, be background noise such as sensor noise, whereas the interference covers other periodic signal not being of interest. Utilizing the periodicity assumption and by exploiting that the desired signal observations across the microphones are just delayed and attenuated version of each other, the signal model can be further specified as

$$\begin{aligned} y_k(n) &= \beta_k s(n - f_{\mathrm{s}}\tau_k) + v_k(n) \\ &= \beta_k \sum_{l=1}^{L} \alpha_l e^{jl\omega_0(n-f_{\mathrm{s}}\tau_k)} + v_k(n), \end{aligned} \tag{2}$$

with L being the number of harmonics, $\alpha_l = A_l e^{j\phi_l}$ being the complex amplitude of the l'th harmonic with A_l and ϕ_l denoting the positive real amplitude and phase, respectively, ω_0 is the fundamental frequency, f_{s} is the sampling frequency, τ_k is the delay of the desired signal from microphone 0 to microphone k, and β_k is the attenuation of the desired signal at sensor k. Note that, by using this model, we have implicitly assumed no reverberation. When the array of microphones is organized in a known way, we can also model the time delay τ_k. For example, if the microphones are organized in a uniform linear array structure, we have that

$$\tau_k = k\frac{d\sin\theta}{c}, \tag{3}$$

where d is the inter microphone spacing, θ is the direction-of-arrival (DOA), and c is the wave propagation speed. That is,

$$y_k(n) = \beta_k \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n} e^{-jl\omega_{\mathrm{s}}k} + v_k(n) \tag{4}$$

with

$$\omega_{\mathrm{s}} = \omega_0 f_{\mathrm{s}} \tau_1 \tag{5}$$

being the so-called spatial frequency. In the remainder of the paper, we assume, for simplicity, that $\beta_p = \beta_q = 1$ for $p \neq q$, which is a reasonable assumption for arrays with closely spaced microphones. When this assumption does not hold, the βs can be estimated using, e.g., the techniques presented in [26]. In practice, N time-consecutive samples from each microphone are used for the estimation of the pitch and DOA. These data can be organized in a matrix like

$$\mathbf{Y}(n) = \begin{bmatrix} y_0(n) & \cdots & y_0(n-N+1) \\ \vdots & \ddots & \vdots \\ y_{K-1}(n) & \cdots & y_{K-1}(n-N+1) \end{bmatrix}. \tag{6}$$

If we consider a subblock of $M \times P$ samples from the above matrix, which is useful for the filter designs to follow later, we can write the signal model on vector form as

$$\begin{aligned} \mathbf{Y}_k(n) &= \begin{bmatrix} y_k(n) & \cdots & y_k(n-M+1) \\ \vdots & \ddots & \vdots \\ y_{k+P-1}(n) & \cdots & y_{k+P-1}(n-M+1) \end{bmatrix} \\ &= \sum_{l=1}^{L} \alpha_l(n)\mathbf{z}_{\mathrm{s}}(l\omega_{\mathrm{s}})\mathbf{z}_{\mathrm{t}}^T(l\omega_0) + \mathbf{V}_k(n), \end{aligned} \tag{7}$$

where $\alpha_l(n) = e^{jl\omega_0 n}$, and

$$\mathbf{z}_{\mathrm{s}}(l\omega_{\mathrm{s}}) = [\,1 \quad e^{-jl\omega_{\mathrm{s}}} \quad \cdots \quad e^{-j(P-1)l\omega_{\mathrm{s}}}\,]^T, \tag{8}$$

$$\mathbf{z}_{\mathrm{t}}(l\omega_0) = [\,1 \quad e^{-jl\omega_0} \quad \cdots \quad e^{-j(M-1)l\omega_0}\,]^T, \tag{9}$$

$$\mathbf{V}_k(n) = \begin{bmatrix} v_k(n) & \cdots & v_k(n-M+1) \\ \vdots & \ddots & \vdots \\ v_{k+P-1}(n) & \cdots & v_{k+P-1}(n-M+1) \end{bmatrix} \tag{10}$$

In the optimal filter designs considered in Section III, it is useful to stack the columns of the subblocks of the observed signal matrix (denoted $\mathrm{vec}\{\cdot\}$), which yields

$$\begin{aligned} \mathbf{y}_k(n) &= \mathrm{vec}\{\mathbf{Y}_k(n)\} \\ &= \sum_{l=1}^{L} \alpha_l(n)\mathbf{z}_l + \mathbf{v}_k(n), \end{aligned} \tag{11}$$

with $\mathbf{v}_k(n) = \mathrm{vec}\{\mathbf{V}_k(n)\}$, and

$$\mathbf{z}_l = \mathrm{vec}\{\mathbf{z}_{\mathrm{s}}(l\omega_{\mathrm{s}})\mathbf{z}_{\mathrm{t}}^T(l\omega_0)\} = \mathbf{z}_{\mathrm{s}}(l\omega_{\mathrm{s}}) \otimes \mathbf{z}_{\mathrm{t}}(l\omega_0), \tag{12}$$

where $\otimes$ denotes the Kronecker product of two vectors or matrices.

A. Motivation for Joint Estimation

Instead of estimating the DOA and pitch jointly, we could estimate those parameters separately with a much lower computational complexity. However, there are a number of significant benefits by conducting the estimation jointly. First of all, in scenarios where multiple periodic sources are present simultaneously, joint estimators may be able to resolve those sources even if either the pitch frequencies or the DOAs of one or more of those sources are similar. This would be impossible if the parameters are estimated separately, since the search is here in only one dimension. Another benefit is a potentially higher estimation accuracy. In [10], it was shown that the asymptotic Cramér-Rao bounds (CRBs) for the DOA and pitch are given by

$$\mathrm{CRB}(\omega_0) \approx \frac{6}{N^3 K}\mathrm{PSNR}^{-1}, \tag{13}$$

$$\begin{aligned} \mathrm{CRB}(\theta) \approx &\left[\left(\frac{c}{\omega_0 f_{\mathrm{s}} d\cos\theta}\right)^2 \frac{6}{NK^3}\right. \\ &\left. + \left(\frac{\tan\theta}{\omega_0}\right)\frac{6}{K^3 N}\right]\mathrm{PSNR}^{-1} \end{aligned} \tag{14}$$

for the scenario described by (11) when $v(n)$ is white noise and $\beta_p \approx \beta_q$ for $p \neq q$, with PSNR denoting the pseudo signal-to-noise ratio. The PSNR is defined as

$$\mathrm{PSNR} = \frac{\sum_{l=1}^{L} l^2 A_l^2}{\sigma_v^2}, \tag{15}$$

and σ_v^2 is the variance of the noise. Close investigation of these expressions reveals the fact that the CRB of the pitch decreases cubically and linearly for increasing N's and K's, respectively. In other words, the pitch estimate can be more accurate when multiple microphone recordings are used. Moreover, we can see that the DOA can be estimated more accurately when taking the harmonic structure of the periodic signal into account as opposed to if the DOA was estimated from, e.g., just the fundamental tone.

Another way of estimating the DOA and pitch is to use a cascaded approach where the DOA is first estimated from the multiple microphone recordings. Then, the signal impinging from this direction is extracted using a beamformer, whereupon the pitch is estimated from the beamformer output. This traditional and cascaded way of estimating the parameters will most likely increase the CRB of the parameter estimated in the second step of this procedure. The cause of this increase, is the linear transformation of the spatio-temporal data introduced by the signal extraction after estimation of the first parameter [10].

III. Spatio-Temporal Filtering Methods

In this section, we present filtering methods for joint estimation of the DOA and pitch from noisy, spatio-temporal, observed data that can be modeled by (11). We assume that we have sampled a signal N times in time and using K sensors in space, which gives us the $K \times N$ data matrix $\mathbf{Y}(n)$ in (6). Then, based on these data, we can design optimal filterbanks or filters for estimating the aforementioned parameters. In all of the presented filtering methods, the idea is to design a filterbank or filter that has minimum output power, while it passes the desired signal undistorted. The joint parameter estimates can then be obtained by maximizing the output power of the so-obtained optimal filters.

A. Optimal Filterbanks

In the filterbank approach, the idea is to design a bank of L FIR filters, where the lth filter should pass the lth harmonic of the desired, periodic signal undistorted. Applying such a bank of FIR filters on a block of the observed signal, we get

$$\begin{aligned} z_k(n) &= \sum_{l=1}^{L}\sum_{m=0}^{M-1}\sum_{p=0}^{P-1} h_{mp}^l y_{k+p}(n-m) \\ &= \sum_{l=1}^{L} \mathbf{1}^T\left[\mathbf{H}_l \circ \mathbf{Y}_k(n)\right]\mathbf{1} \\ &= \sum_{l=1}^{L} \mathbf{h}_l^H \mathbf{y}_k(n), \end{aligned} \tag{16}$$

where $\circ$ denotes the Hadamard product, M and P are the temporal and spatial filter lengths, respectively, h_{mp}^l is the (m,p)th coefficient of the lth filter in the filterbank, $\mathbf{h}_l = \mathrm{vec}\{\mathbf{H}_l\}$, $\mathbf{y}_k(n) = \mathrm{vec}\{\mathbf{Y}_k(n)\}$, $\mathbf{1}$ is a column vector of ones, and

$$\mathbf{H}_l = \begin{bmatrix} h_{00}^l & \cdots & h_{(M-1)0}^l \\ \vdots & \ddots & \vdots \\ h_{0(P-1)}^l & \cdots & h_{(M-1)(P-1)}^l \end{bmatrix}. \tag{17}$$

Then, we can design a filterbank where the sum of the output powers from the individual filters is minimized, while the lth filter passes the lth harmonic undistorted and cancels out the other harmonics. The sum of the output powers of the filters is given by

$$\sum_{l=1}^{L} \mathrm{E}\left[\left|\mathbf{h}_l^H \mathbf{y}_k(n)\right|^2\right] = \sum_{l=1}^{L} \mathbf{h}_l^H \mathbf{R}_{\mathbf{y}} \mathbf{h}_l, \tag{18}$$

where $\mathbf{R}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}_k(n)\mathbf{y}_k^H(n)]$ is the covariance matrix of $\mathbf{y}_k(n)$. From (11) and (18), it is clear that the aforementioned

design goal can be achieved by solving the following optimization problem:

$$\min_{\mathbf{H}} \mathrm{Tr}[\mathbf{H}^H \mathbf{R}_{\mathbf{y}} \mathbf{H}] \quad \text{s.t.} \quad \mathbf{H}^H \mathbf{Z} = \mathbf{I}, \tag{19}$$

with

$$\mathbf{Z} = [\, \mathbf{z}_1 \quad \mathbf{z}_2 \quad \cdots \quad \mathbf{z}_L \,]. \tag{20}$$

The well known solution to this second order optimization problem can obtained using Lagrange multipliers, and it is given by

$$\mathbf{H}_{\mathrm{opt}} = \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z} \big(\mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z} \big)^{-1}. \tag{21}$$

Note that for $\mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z}$ to be invertible, we must require that $M \geq L$, and this is also the case for the other optimal filter designs proposed in the remainder of the section. Interestingly, it can be shown that a filter identical to the one in (21) can be designed by minimizing the sum of the powers of the noise at the output of all the filters [27], which gives [18]

$$\mathbf{H}_{\mathrm{opt}} = \mathbf{R}_{\mathbf{v}}^{-1} \mathbf{Z} \big(\mathbf{Z}^H \mathbf{R}_{\mathbf{v}}^{-1} \mathbf{Z} \big)^{-1}. \tag{22}$$

This fact can be exploited to achieve some computationally more efficient filter designs. If we, for a moment, assume that the noise is white such that $\mathbf{R}_{\mathbf{v}} = \sigma_v^2 \mathbf{I}$, where σ_v^2 is the variance of the noise, we get that

$$\mathbf{H}_{\mathrm{wn}} = \mathbf{Z} \big(\mathbf{Z}^H \mathbf{Z} \big)^{-1}. \tag{23}$$

This filter will, of course, only be optimal with respect to the aforementioned design criteria when the noise is indeed white, but it may still be useful even in other noise settings due to its simplicity. Finally, we can achieve an approximative filter design by exploiting that [9]

$$\lim_{MP \to \infty} \frac{1}{MP} \mathbf{Z}^H \mathbf{Z} = \mathbf{I}. \tag{24}$$

In this case, the optimal filter for the white noise scenario becomes

$$\mathbf{H}_{\mathrm{awn}} = \frac{1}{MP} \mathbf{Z}. \tag{25}$$

This approximative filter design can be interpreted as a filterbank of spatio-temporal, periodogram-based filters [18], and it can be applied efficiently in practice using FFTs.

Using either of the aforementioned filter designs, the fundamental frequency and the DOA can then be estimated jointly. This is achieved by maximizing the sum of the output powers of the filters in these filterbanks over sets of candidate fundamental frequencies and DOAs, i.e.,

$$\{\widehat{\omega}_0, \widehat{\theta}\} = \arg \max_{\{\omega_0, \theta\} \in \Omega \times \Theta} \mathrm{Tr} \left[\mathbf{H}^H \mathbf{R}_{\mathbf{y}} \mathbf{H} \right], \tag{26}$$

where Ω and Θ are the sets of candidate fundamental frequencies and DOAs, respectively. We note that, in practice, $\mathbf{R}_{\mathbf{y}}$ is most likely not known and therefore has to be estimated. Moreover, it is worth mentioning that, while the above estimator is only for estimating the pitch and DOA of a single source, the estimator can be used in the iterative RELAX algorithm in [28] to estimate the parameters of multiple sources.

B. Optimal Single Filters

An alternative filtering approach to joint fundamental frequency and DOA estimation is the single filter approach. In this approach, the idea is to apply a single FIR filter on a block of the observed signal, yielding the output:

$$\begin{aligned} z_k(n) &= \sum_{m=0}^{M-1} \sum_{p=0}^{P-1} h_{mp} y_{k+p}(n-m) \\ &= \mathbf{1}^T [\mathbf{H} \circ \mathbf{Y}_k(n)] \mathbf{1} \\ &= \mathbf{h}^H \mathbf{y}_k(n), \end{aligned} \tag{27}$$

where $\mathbf{H}$ is defined similarly to $\mathbf{H}_l$, i.e.,

$$\mathbf{H} = \begin{bmatrix} h_{00} & \cdots & h_{(M-1)0} \\ \vdots & \ddots & \vdots \\ h_{0(P-1)} & \cdots & h_{(M-1)(P-1)} \end{bmatrix}, \tag{28}$$

and $\mathbf{h} = \mathrm{vec}\{\mathbf{H}\}$. We then want to design a single filter that passes all of the harmonics undistorted while the output power of the filter is minimized. The output power of the single filter is given by

$$\mathrm{E} \left[\left| \mathbf{h}^H \mathbf{y}_k(n) \right|^2 \right]. \tag{29}$$

That is, a solution to the above filter design problem is, from (11), clearly achieved by solving:

$$\begin{aligned} \min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_{\mathbf{y}} \mathbf{h} \quad &\text{s.t.} \quad \mathbf{h}^H \mathbf{z}_l = 1, \\ &\text{for} \quad l = 1, \ldots, L. \end{aligned} \tag{30}$$

Like in the filterbank approach, the solution to this optimization problem can be obtained using Lagrange multipliers and is given by

$$\mathbf{h}_{\mathrm{opt}} = \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z} \big(\mathbf{Z}^H \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{Z} \big)^{-1} \mathbf{1}. \tag{31}$$

The same filter is obtained if we minimize the power of the noise after filtering under the same constraints, in which case the optimal filter is given by [27]

$$\begin{aligned} \mathbf{h}_{\mathrm{opt}} &= \mathbf{R}_{\mathbf{v}}^{-1} \mathbf{Z} \big(\mathbf{Z}^H \mathbf{R}_{\mathbf{v}}^{-1} \mathbf{Z} \big)^{-1} \mathbf{1} \\ &= \mathbf{H}_{\mathrm{opt}} \mathbf{1}. \end{aligned} \tag{32}$$

If we again assume that the noise is white, with $\mathbf{R}_{\mathbf{v}} = \sigma_v^2 \mathbf{I}$, we get that

$$\mathbf{h}_{\mathrm{wn}} = \mathbf{Z} \big(\mathbf{Z}^H \mathbf{Z} \big)^{-1} \mathbf{1} = \mathbf{H}_{\mathrm{wn}} \mathbf{1}. \tag{33}$$

Then, by applying the approximation in (24), we can obtain an approximative single filter design as

$$\mathbf{h}_{\mathrm{awn}} = \frac{1}{MP} \mathbf{Z} \mathbf{1} = \mathbf{H}_{\mathrm{awn}} \mathbf{1}. \tag{34}$$

This filter can be seen as a sum of spatio-temporal, periodogram-based filters.

In the single filter approach, the fundamental frequency and DOA are then jointly estimated by simply maximizing the output power of either of the filter designs proposed above over the sets of candidate fundamental frequencies Ω and DOAs Θ.

TABLE I
COST FUNCTIONS INVOLVED IN THE ESTIMATORS OBTAINED USING THE DIFFERENT FILTERING APPROACHES AND DESIGN TOPOLOGIES

	Filterbank	**Single Filter**
Optimal	$\text{Tr}\left[\left(\mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\right]$	$\mathbf{1}^T\left(\mathbf{Z}^H\mathbf{R}_\mathbf{y}^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}$
White Noise	$\text{Tr}\left[\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{R}_\mathbf{y}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\right]$	$\mathbf{1}^T\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{R}_\mathbf{y}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{1}$
Approx.	$\frac{1}{M^2P^2}\text{Tr}\left[\mathbf{Z}^H\mathbf{R}_\mathbf{y}\mathbf{Z}\right]$	$\frac{1}{M^2P^2}\mathbf{1}^T\mathbf{Z}^H\mathbf{R}_\mathbf{y}\mathbf{Z}\mathbf{1}$

Mathematically speaking, the joint estimates obtained using the single filter approach can be put as

$$\{\widehat{\omega}_0, \widehat{\theta}\} = \arg \max_{\{\omega_0,\theta\}\in\Omega\times\Theta} \mathbf{h}^H\mathbf{R}_\mathbf{y}\mathbf{h}. \tag{35}$$

In Table I, an overview of the estimators obtained using the two different filtering approaches and the different filter design topologies is found. These have been found by first inserting the expressions in (21), (23), and (25) in (26), and then inserting (31), (33), and (34) in (35). We note that, except for a scaling factor of $\frac{1}{M^2P^2}$, the cost function for the approximative filterbank approach resembles that of the approximative NLS estimator in [10]. Furthermore, as mentioned in Section III-A, the above estimator can be used in an iterative algorithm to estimate the parameters of multiple sources [28].

When the assumptions on white Gaussian noise and large sample sizes do not hold, the white noise and approximative filtering methods will consequently not be optimal, and most often yield less accurate estimates compared to the optimal filtering methods as we also discovered in the experiments in Section IV. The approximative filtering methods, however, are computationally simpler compared to the optimal filtering methods by not requiring any inversions. That is, the filter design can be chosen to achieve a certain tradeoff between computational complexity and estimation accuracy.

C. Estimation of the Covariance Matrix

In the estimators presented in this section, knowledge about the covariance matrix $\mathbf{R}_\mathbf{y}$ is needed. This covariance matrix is obviously not known in most practical scenarios, so we need to replace it by an estimate. One possible estimate is the outer product estimate, which is commonly used, e.g., in single-channel, fundamental frequency estimation. In the multichannel, spatio-temporal case, the outer product estimate of $\mathbf{R}_\mathbf{y}$ is given by

$$\widehat{\mathbf{R}}_\mathbf{y} = \sum_{k=0}^{K-P}\sum_{m=0}^{N-M} \frac{\mathbf{y}_k(n-m)\mathbf{y}_k^H(n-m)}{(K-P+1)(N-M+1)}. \tag{36}$$

The optimal estimators for the general noise case (see Table I require the covariance matrix of the observed signal to be inverted. To ensure that $\widehat{\mathbf{R}}_\mathbf{y}$ is invertible, we must require it to be full-rank, i.e.,

$$(K-P+1)(N-M+1) \geq MP \tag{37}$$

needs to be fulfilled. Typically, $K \ll N$ and P is desired to be as large as possible to attain a reasonable spatial resolution. If we, for example, choose $P = K$ we have that

$$M \leq \frac{N+1}{K+1}. \tag{38}$$

As a result of that, M may need to be very small or a large amount of temporal samples N is needed if K is relatively large.

Alternatively, to circumvent this issue, an iterative adaptive approach (IAA) [22], [23] on the estimation of the covariance matrix can be taken. First, let the amplitude of a spatio-temporal frequency component of interest be denoted by $\alpha_{\gamma',\psi'}$, where γ' is a frequency index, and ψ' is a direction index corresponding to the DOA. Then, using the covariance matrix model, the noise covariance matrix can be approximated as

$$\begin{aligned}\mathbf{Q}_{\gamma',\psi'} &\approx \sum_{\gamma=1}^{\Gamma}\sum_{\psi=1}^{\Psi} |\alpha_{\gamma,\psi}|^2\mathbf{z}_{\gamma,\psi}\mathbf{z}_{\gamma,\psi}^H - |\alpha_{\gamma',\psi'}|^2\mathbf{z}_{\gamma',\psi'}\mathbf{z}_{\gamma',\psi'}^H \\ &= \widetilde{\mathbf{R}}_\mathbf{y} - |\alpha_{\gamma',\psi'}|^2\mathbf{z}_{\gamma',\psi'}\mathbf{z}_{\gamma',\psi'}^H,\end{aligned} \tag{39}$$

where γ and ψ denote frequency and direction indices, respectively, Γ is the number of frequency grid points utilized in the IAA, Ψ is the number direction grid points utilized in the IAA,

$$\widetilde{\mathbf{R}}_\mathbf{y} = \sum_{\gamma=1}^{\Gamma}\sum_{\psi=1}^{\Psi} |\alpha_{\gamma,\psi}|^2\mathbf{z}_{\gamma,\psi}\mathbf{z}_{\gamma,\psi}^H, \tag{40}$$

is an estimate of the observed signal covariance matrix, and

$$\mathbf{z}_{\gamma,\psi} = \mathbf{z}_\text{s}(\psi,\gamma)\otimes\mathbf{z}_\text{t}(\gamma), \tag{41}$$

$$\mathbf{z}_\text{t}(\gamma) = \begin{bmatrix}1 & e^{-j\omega_\gamma} & \cdots & e^{-j(N-1)\omega_\gamma}\end{bmatrix}^T, \tag{42}$$

$$\mathbf{z}_\text{s}(\psi,\gamma) = \begin{bmatrix}1 & e^{-j\omega_{\text{s},\psi,\gamma}} & \cdots & e^{-j(K-1)\omega_{\text{s},\psi,\gamma}}\end{bmatrix}^T, \tag{43}$$

with ω_γ denoting the frequency corresponding to the γ'th grid point, and $\omega_{\text{s},\psi,\gamma}$ denoting the spatial frequency corresponding to the grid points ψ and γ, i.e.,

$$\omega_{\text{s},\psi,\gamma} = \omega_\gamma f_\text{s}\frac{d\sin\theta_\psi}{c}. \tag{44}$$

In (44), θ_ψ is the DOA corresponding to the ψ'th grid point.

The IAA is then used to obtain an estimate of the amplitude $\alpha_{\gamma',\psi'}$ by minimizing a weighted least-squares (WLS) cost function J_WLS given by

$$J_\text{WLS} = \left[\mathbf{y}(n) - \alpha_{\gamma',\psi'}\mathbf{z}_{\gamma',\psi'}\right]^H\mathbf{Q}_{\gamma',\psi'}\left[\mathbf{y}(n) - \alpha_{\gamma',\psi'}\mathbf{z}_{\gamma',\psi'}\right], \tag{45}$$

TABLE II
IAA FOR SPATIO-TEMPORAL SPECTRUM AND COVARIANCE ESTIMATION

initialization

$$\hat{\alpha}_{\gamma,\psi} = \frac{\mathbf{z}_{\gamma,\psi}^H \mathbf{y}(n)}{\mathbf{z}_{\gamma,\psi}^H \mathbf{z}_{\gamma,\psi}}, \quad \gamma = 1,\ldots,\Gamma, \quad \psi = 1,\ldots,\Psi$$

repeat

$$\widetilde{\mathbf{R}}_{\mathbf{y}} = \sum_{\gamma=1}^{\Gamma}\sum_{\psi=1}^{\Psi} |\alpha_{\gamma,\psi}|^2 \mathbf{z}_{\gamma,\psi}\mathbf{z}_{\gamma,\psi}^H$$

$$\hat{\alpha}_{\gamma,\psi} = \frac{\mathbf{z}_{\gamma,\psi}^H \widetilde{\mathbf{R}}_{\mathbf{y}}^{-1} \mathbf{y}(n)}{\mathbf{z}_{\gamma,\psi}^H \widetilde{\mathbf{R}}_{\mathbf{y}}^{-1} \mathbf{z}_{\gamma,\psi}}, \quad \gamma = 1,\ldots,\Gamma, \quad \psi = 1,\ldots,\Psi$$

until (convergence)

with $\mathbf{y}(n) = \mathrm{vec}\{\mathbf{Y}(n)\}$. Minimizing the cost function with respect to the unknown amplitude $\alpha_{\gamma',\psi'}$ yields the following closed-form estimate

$$\hat{\alpha}_{\gamma',\psi'} = \frac{\mathbf{z}_{\gamma',\psi'}^H \mathbf{Q}_{\gamma',\psi'}^{-1} \mathbf{y}(n)}{\mathbf{z}_{\gamma',\psi'}^H \mathbf{Q}_{\gamma',\psi'}^{-1} \mathbf{z}_{\gamma',\psi'}}. \tag{46}$$

Using the matrix inversion lemma on (39), it can be shown that the amplitude estimate is equivalently found from

$$\hat{\alpha}_{\gamma',\psi'} = \frac{\mathbf{z}_{\gamma',\psi'}^H \widetilde{\mathbf{R}}_{\mathbf{y}}^{-1} \mathbf{y}(n)}{\mathbf{z}_{\gamma',\psi'}^H \widetilde{\mathbf{R}}_{\mathbf{y}}^{-1} \mathbf{z}_{\gamma',\psi'}}. \tag{47}$$

This expression is preferred over (46) as the covariance matrix estimate $\widetilde{\mathbf{R}}_{\mathbf{y}}$ needs to be formed only once, while $\mathbf{Q}_{\gamma',\psi'}$ needs to be updated per frequency and direction grid point. We note that the amplitude estimate depends on the estimate of the covariance matrix and vice versa, so these are estimated by iterating between (40) and (47), hence the method is termed the IAA. While the IAA has historically been used for amplitude spectrum estimation, we here utilize it for estimation of the covariance matrix of the observed signal herein. As opposed to the sample covariance matrix estimate, this estimate is formed from a single observation, $\mathbf{y}(n)$, while also being full-rank. This enables us to choose $M = N$ and $P = K$, but of course it is computationally more complex to obtain this estimate than the sample covariance matrix estimate. The algorithm is summarized in Table II. As it can be seen, the algorithm is initialized with $\widetilde{\mathbf{R}}_{\mathbf{y}} = \mathbf{I}$. Typically, 10-15 iterations is sufficient to achieve convergence in practice.

IV. EXPERIMENTAL RESULTS

We now proceed with an experimental evaluation of the proposed filter designs. The evaluation is split into three parts: 1) a qualitative comparison of the proposed filters, 2) a thorough statistical evaluation of the proposed filters through Monte-Carlo simulations including comparison with state of the art, and 3) qualitative evaluation of the filters on a real-life signal. First, we compare the cost functions of the optimal filterbank and single filter when using both the sample and IAA-based covariance matrix estimates. For this experiment, we used a synthetic, periodic signal with $L = 4$ harmonics with unit amplitudes, $\theta = -50°$, $f_0 = 200$ Hz, $f_s = 4$ kHz, and white noise was added to each sensor signal at an SNR of

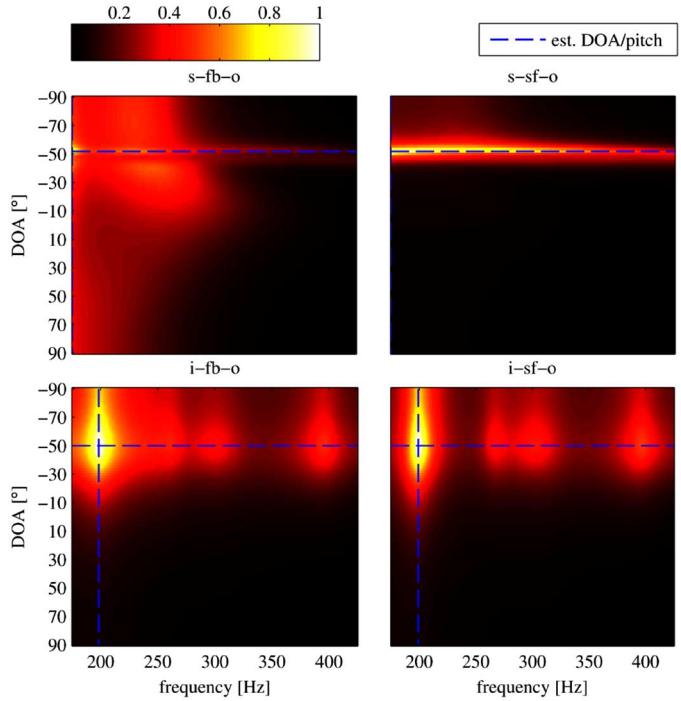

Fig. 1. Plots of the cost functions for the optimal (left) filterbank and (right) single filter methods implemented using the (top) sample and (bottom) IAA-based covariance matrix estimates when applied on a synthetic, multichannel, periodic signal for $N = 20$.

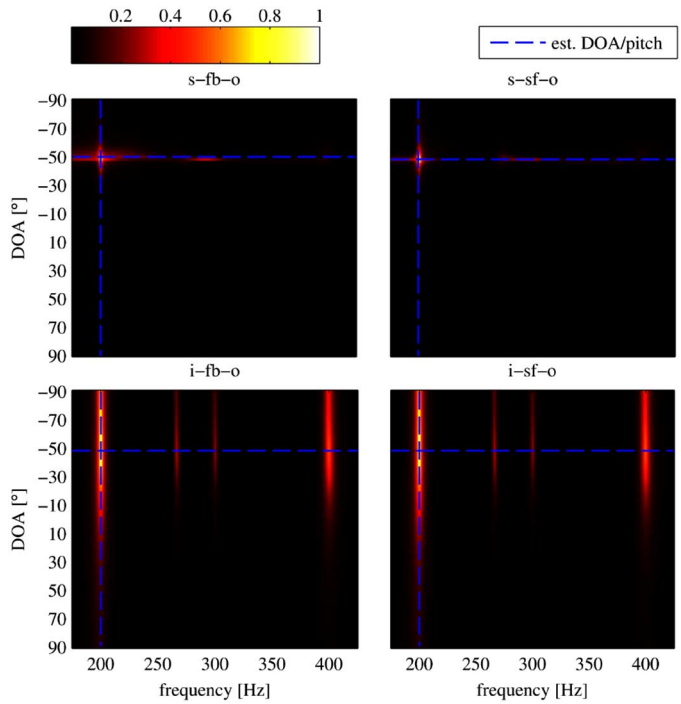

Fig. 2. Plots of the cost functions for the optimal (left) filterbank and (right) single filter methods implemented using the (top) sample and (bottom) IAA-based covariance matrix estimates when applied on a synthetic, multichannel, periodic signal for $N = 60$.

30 dB. The other parameters of interest in the simulation were chosen as follows: $K = 3$, $P = 3$, $N = 20$, $M = \lfloor (N+1)/(K+1) \rfloor$, $\Gamma = 512$, $\Psi = 128$, 10 iterations was used to obtain the IAA estimate, $c = 343$ m/s, and $d = 0.04$ m. Using this setup, we then evaluated the cost functions of the optimal filters

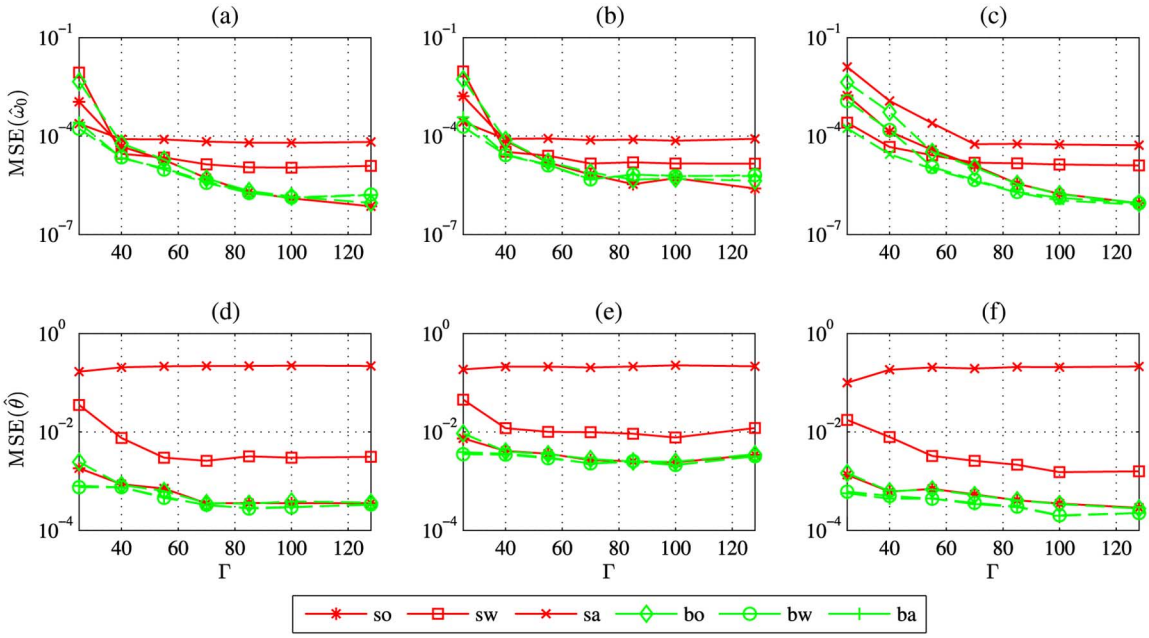

Fig. 3. MSEs of pitch and DOA estimates for different Γ's in scenarios with (a,d) $N = 20$ and an SNR of 30 dB, (b,e) $N = 20$ and an SNR of 20 dB, and (c,f) $N = 25$ and an SNR of 30 dB for the proposed methods.

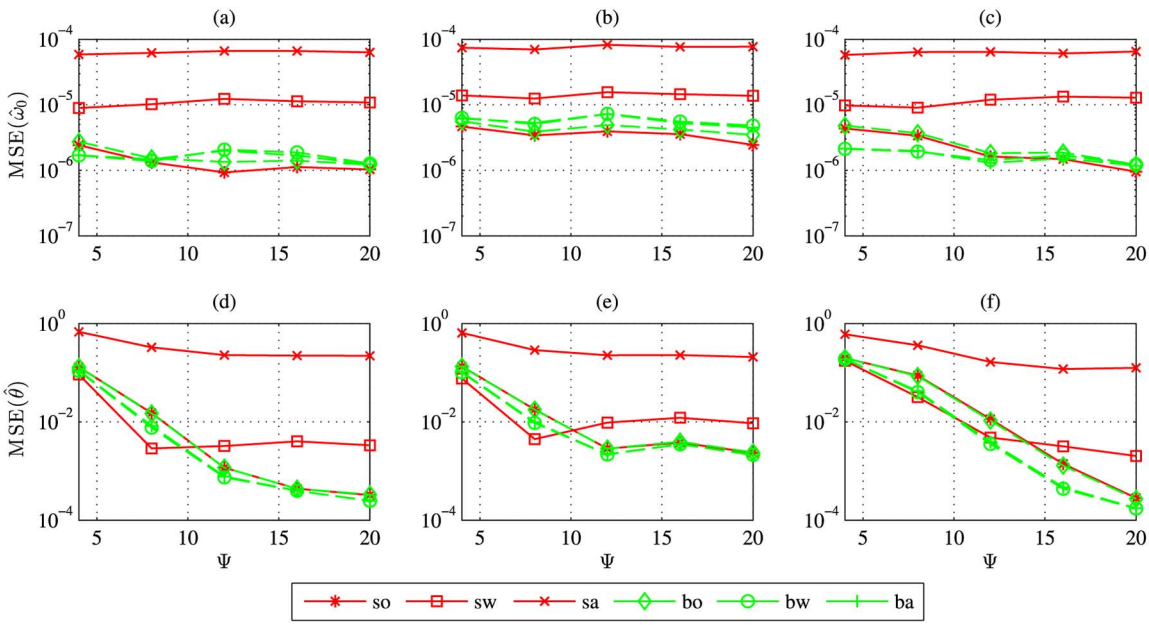

Fig. 4. MSEs of pitch and DOA estimates for different Ψ's in scenarios with (a,d) $K = 2$ and an SNR of 30 dB, (b,e) $K = 2$ and an SNR of 20 dB, and (c,f) $K = 3$ and an SNR of 30 dB for the proposed methods.

in Table I for different candidate pitch frequencies and DOAs when using the sample and IAA-based covariance matrix estimates, and the results are depicted in Fig. 1. From the figures, we can see that none of the methods show a distinct peak at the true DOA and pitch when the sample covariance matrix estimate is used. This is opposed to when using the IAA-based covariance matrix estimate, in which case both optimal filtering methods each yield a distinct, maximum peak near the true parameters. This indicates that for low numbers of samples, the IAA-based covariance matrix estimate should be used. Moreover, it supports the practicability of optimal filtering with the IAA despite its computational complexity, since small sample sizes are generally preferred when the signal of interest is nonstationary, violating the stationarity assumption in (4). This is often the case in practice, e.g., when processing speech signals. The same experiment was conducted for $N = 80$ resulting in the cost functions in Fig. 2. For this sample length, both optimal filtering methods seem to provide a good estimate of the DOA and pitch for both covariance matrix estimates. However, the sample covariance matrix estimate seems to give the best resolution in this case due to narrower peaks around the true parameters, with the optimal single filter having the narrowest peak. This indicates that, for longer sample sizes, the sample covariance matrix estimate may be preferred.

In the another series of experiments, we evaluated the statistical performance in terms of mean squared errors (MSEs) of the proposed estimators implemented using the IAA-based covariance matrix estimate (since relatively small sample sizes

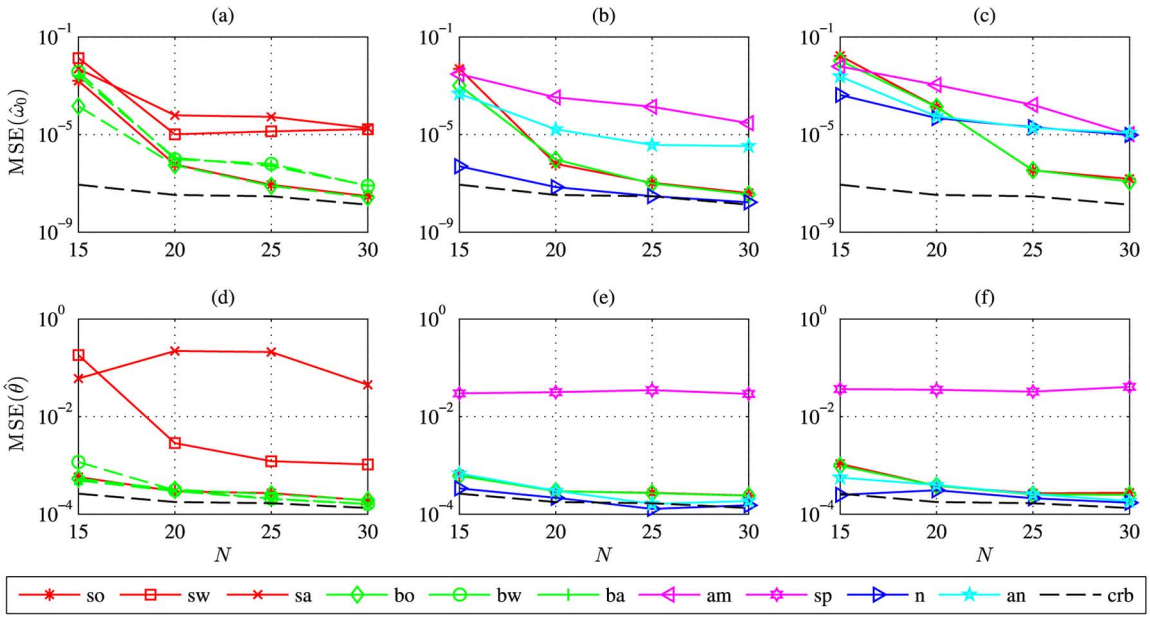

Fig. 5. MSEs of pitch and DOA estimates for different N's in scenarios with (a,b,d,e) white noise, and (c,f) white noise and an interfering source for the proposed and state-of-the-art methods.

are considered) through Monte Carlo simulations. In all these experiments, 100 Monte Carlo simulations were conducted for each parameter setting, and, in each simulation, the noise and the phases of the harmonics were randomized. The MSEs of the pitch and DOA estimates (MSE_{ω_0} and MSE_{θ}), respectively) obtained from these simulations were calculated as

$$\text{MSE}(\widehat{\omega}_0) = \frac{1}{Q}\sum_{q=1}^{Q}(\omega_{0,q} - \widehat{\omega}_{0,q})^2, \tag{48}$$

$$\text{MSE}(\widehat{\theta}) = \frac{1}{Q}\sum_{q=1}^{Q}(\theta_q - \widehat{\theta}_q)^2, \tag{49}$$

where Q is the number of Monte Carlo simulations, q is the simulation number, $\omega_{0,q}$ and θ_q are the true pitch and DOA in simulation q, and $\widehat{(\cdot)}$ denotes an estimate of a parameter.

Moreover, a synthetic, multichannel periodic signal was used in every simulation with $L = 4$ harmonics with unit amplitudes, and, in each simulation, the pitch and DOA were sampled from $\mathcal{U}(250\text{ Hz}, 300\text{ Hz})$ and $\mathcal{U}(15°, 35°)$, respectively, where $\mathcal{U}(a, b)$ denotes the continuous uniform distribution in the interval $[a; b]$. The methods evaluated in these experiments are the optimal, white noise, and approximate filterbank ('bo', 'bw', and 'ba') and single filter ('so', 'sw', and 'sa') methods, the multichannel pitch estimator ('am') in [29], the steered response power method with phase transform ('sp') [30], and the exact and asymptotic nonlinear least squares (NLS) methods ('n' and 'an') in [10]. First, the performance of the proposed methods was evaluated for different Γ's in scenarios with 1) $N = 20$ and an SNR of 30 dB, 2) $N = 20$ and an SNR of 20 dB, and 3) $N = 25$ and an SNR of 30 dB. The other simulation parameters were: $f_s = 4$ KHz, $c = 343$ m/s, $K = 2$, $d = 0.04$ m, and $\Psi = 64$. The results from this series of simulations are depicted in Fig. 3. From this figure, we make two important observations: first, the performances of the 'sw' and 'sa' methods are generally worse than those of the other proposed methods. Moreover, the results indicate that the higher the SNR and number of samples N, the more frequency grid points Γ is needed in the IAA-based covariance matrix estimation to achieve the highest possible performance. A similar series of simulations were conducted where the performance of the proposed methods were evaluated for different Ψ's. In these experiments, three scenarios were considered: 1) $K = 2$ and an SNR of 30 dB, 2) $K = 2$ and an SNR of 20 dB, and 3) $K = 3$ and an SNR of 30 dB. The other simulation parameters were the same as in the previous series of simulations except that $N = 20$ and $\Gamma = 100$, and the results are provided in Fig. 4. As in the previous series of simulations, we observe that the higher the SNR and number of sensors, the more spatial frequency grid points is needed in the IAA-based covariance matrix estimation to achieve the maximum possible performance.

Then, we conducted other series of simulations where the performance of the proposed methods were also compared with the state-of-the-art methods mentioned before. In the first of these evaluations, the performance was measured for different N's in two scenarios: 1) a scenario where the periodic signal was added with white noise at an SNR of 30 dB, and 2) a scenario with both white noise and an interfering source added where the SNR was 30 dB, and the interfering source was a single sinusoid with unit amplitude and random phase. The interfering sinusoid had the same DOA as the desired signal, but a frequency equal to $f_i = f_0 + 60$ Hz. Otherwise, the simulation parameters were chosen as in the previous Monte-Carlo simulations except that $\Gamma = 512$, $\Psi = 64$, and $K = 2$. The results are found in Fig. 5. First of all, we observe that the proposed 'so', 'bo', 'bw', and 'ba' methods all yield similar performance, and that they outperform the 'sw' and 'sa' methods for the whole range of N's. In the comparison with state of the art, we see that the 'n' method generally has the best performance in the white noise only scenario. However, for higher N's (≥ 25), there is not much difference between the proposed optimal filtering methods and the

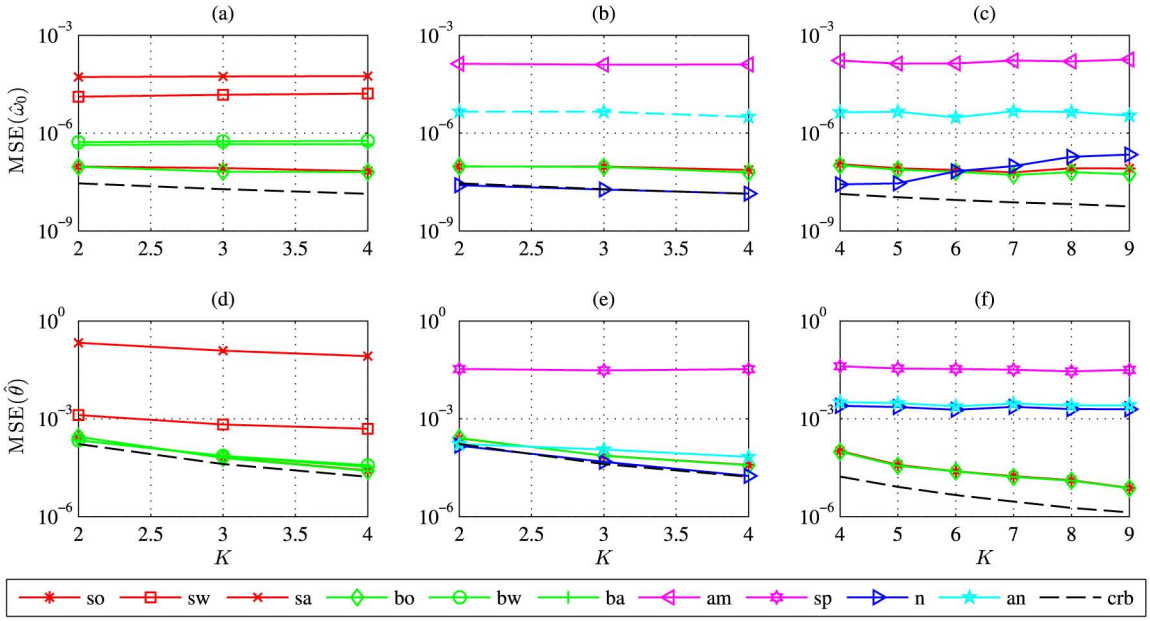

Fig. 6. MSEs of pitch and DOA estimates for different K's in scenarios with (a,b,d,e) white noise, and (c,f) white noise and an interfering source for the proposed and state-of-the-art methods.

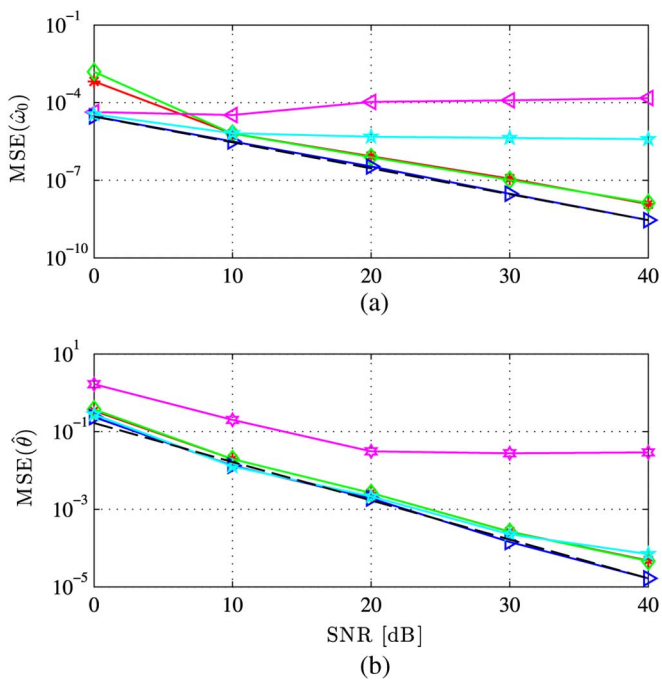

Fig. 7. MSEs of pitch and DOA estimates for different SNRs in a scenario with white noise for the proposed and state-of-the-art methods. The labels for the plot are similar to those in Fig. 5.

'n' method and, for $N \geq 20$, the proposed methods clearly outperform the 'an' and 'am' methods for pitch estimation and the 'sp' method for DOA estimation. Finally, in the scenario with an interfering source, the proposed optimal filters clearly outperform all other methods for pitch estimation in the range $25 \leq N \leq 30$, while they are only slightly worse than the 'n' and 'an' methods for DOA estimation in general. Similarly, we also evaluated the performance for different K's. Again, a scenario with white noise and a scenario with white noise and an additional interfering sinusoid with unit amplitude were considered. In this evaluation, however, the interfering source had the

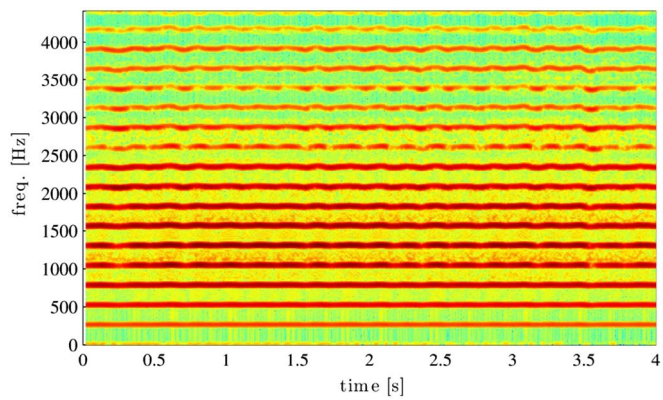

Fig. 8. Plot of the spectrogram of a single-channel trumpet signal with vibrato.

same frequency as the pitch of the harmonic signal, while its DOA was $\theta_i = \theta - 80°$. The IAA grid size parameters were $\Gamma = 256$ and $\Psi = 128$, the number of temporal samples was $N = 25$, and otherwise the simulation parameters were the same as in the previous Monte-Carlo simulations. The results from this experiment are depicted in Fig. 6. For pitch estimation, the 'bw' and 'ba' generally yield the best performance of the proposed methods, followed closely by the 'bo', 'so', 'sa', and 'sw' methods in this order. For DOA estimation, the 'so', 'bo', 'bw', and 'ba' methods yield similar performance and outperform the 'sw' and 'sa' methods. In comparison with state of the art in the white noise scenario, we see that the 'so' and 'bo' methods have similar performance to the 'an' method for pitch estimation and that they outperform the 'am' method. The 'n' method generally yields the most accurate pitch estimates, though. The same observations are also valid for DOA estimation for the 'so', 'bo', 'an' and 'n' methods, whereas the 'sp' method shows a much worse performance. In the scenario where an additional interfering sinusoid is added, the proposed 'so' and 'bo' outperform the 'an' and 'am' methods for all K's, wheres the 'n' method shows better performance for low K's due to bias and worse

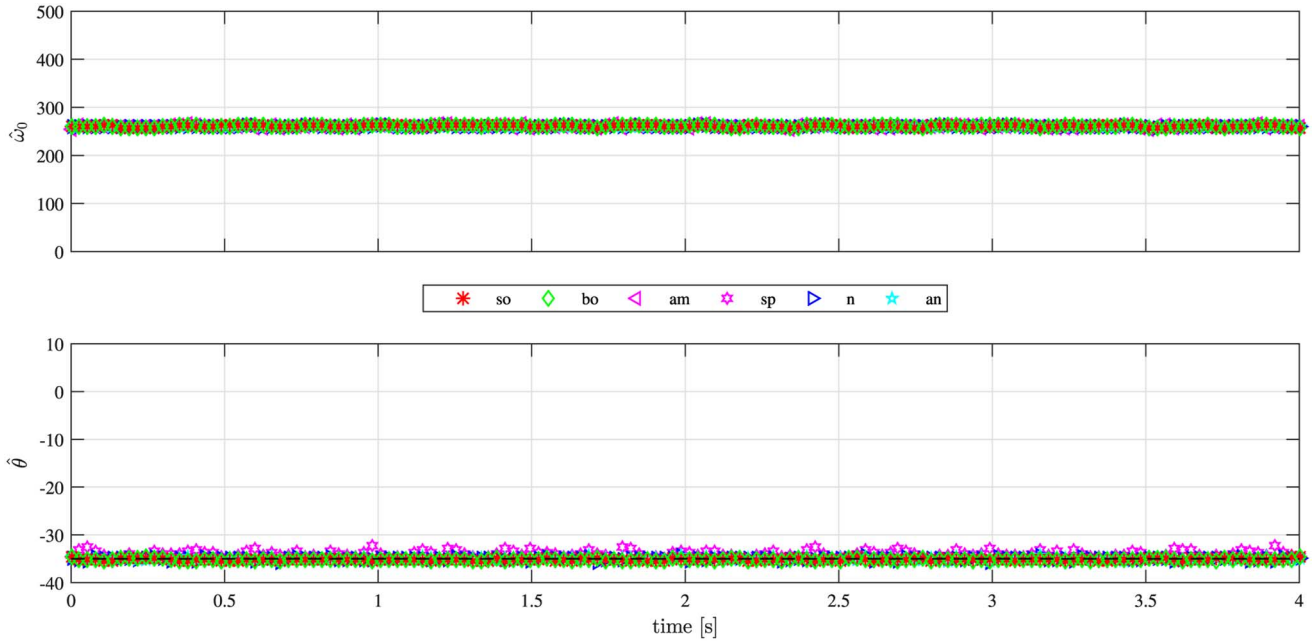

Fig. 9. Plots of (top) pitch and (bottom) DOA estimates obtained from the spatially resynthesized, trumpet signal in a scenario with no reverberation.

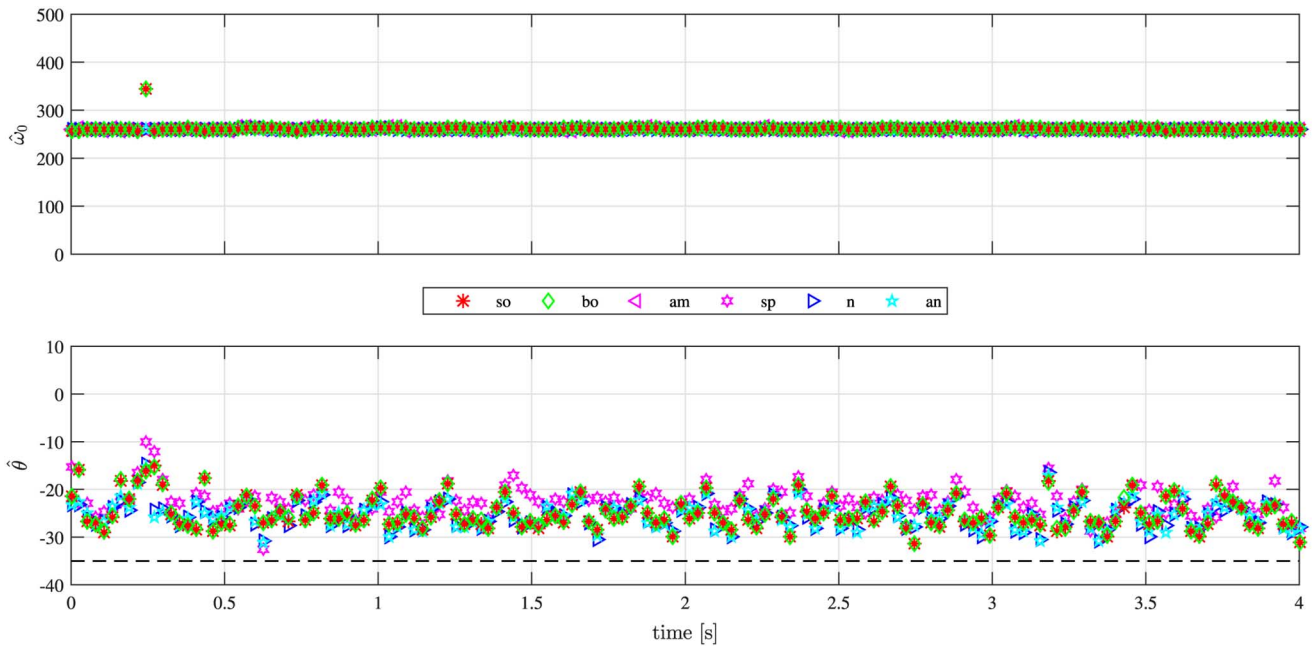

Fig. 10. Plots of (top) pitch and (bottom) DOA estimates obtained from the spatially resynthesized, trumpet signal in a scenario with reverberation and a reverberation time of $T_{60} = 0.5$ s.

performance for higher K's. For DOA estimation the proposed optimal filtering methods clearly outperforms all other methods in the comparison.

In the last series of Monte-Carlo simulations, the performances were measured for different SNRs in a scenario with white noise only. The setup for these simulations was: $N = 25$, $K = 2$, $\Gamma = 512$, $\Psi = 50$, and the remaining parameters were setup as in the previous Monte-Carlo simulations. We see, from the results in Fig. 7, that the 'n' method has the best performance as expected for both DOA and pitch estimation in all scenarios, however, the difference in terms of DOA estimation performance between the 'n', 'an', 'so', and 'bo' methods is negligible for low SNRs ($\leq$ 30 dB). In terms of pitch estimation, the proposed methods outperform the 'am' and 'an' methods for SNR $\geq$ 20 dB, and, for DOA estimation, the 'sp' method is outperformed in all scenarios.

A final evaluation of the proposed filtering methods was conducted on a real-life signal. The signal used in this experiment was a 4 seconds long, single-channel trumpet signal with vibrato. The spectrogram of the signal is shown in Fig. 8, and it can be seen that it has a pitch fluctuating around ≈ 260. Based on the spectrogram, we chose a fixed model order for the experiment of $L = 5$. To obtain a multichannel signal, the signal was resynthesized spatially, using an online available room impulse response (RIR) generator [31]. The RIR generator was set up as follows: $c = 343$ m/s, $f_s = 8,820$ Hz, the micro-

phones of a ULA with 5 sensors was located at $\langle 2 + d[k - (K - 1)/2]\rangle \mathrm{m} \times 0.5\ \mathrm{m} \times 1.5\ \mathrm{m}$, for $k = 1, \ldots, 5$, $d = 0.04$ m, the source was located at $\theta = -35^\circ$ at a distance of 3 m from the center of the array, the room dimensions were $4\ \mathrm{m} \times 4\ \mathrm{m} \times 3\ \mathrm{m}$, the length of the RIRs was 2048, the microphones had cardioid responses with orientation $(90^\circ, 0^\circ)[(\mathrm{azimuth}, \mathrm{elevation})]$, and the reflection order was 0. Then, we generated the multichannel, real data using this setup, and applied the proposed optimal filtering methods and the state of the art methods on time-consecutive frame of length $N = 40$ of the signal. The methods were implemented with $\Gamma = 128$, $\Psi = 64$, and, in the 'sp' method, we used an FFT length of 256 and integrated over frequencies in the interval $[150\ \mathrm{Hz}, f_s/2]$. From this experiment, we obtained the results depicted in Fig. 9. The results show that all methods yield pitch estimates close to the true pitch by comparing the estimates with the spectrogram of the trumpet signal. Moreover, we see that the proposed 'so' and 'bo' methods along with the 'an' and 'n' methods obtain DOA estimates closer to the true DOA than the 'sp' method at most time instances. Subsequently, a similar experiment was carried out where reverberation was added, i.e., the same simulation setup was used except that the reflection order was set to -1 (maximum), and the reverberation time was 0.5 s. With this setup, we obtained the results in Fig. 10. Again, all methods seem to provide pitch estimates close to the true pitch. The DOA estimates obtained using all methods are less accurate and biased in this scenario. In general, the proposed 'so' and 'bo' methods seem to perform similar to the 'n', 'an' methods in terms of accuracy, whereas the 'sp' method is generally outperformed.

V. Conclusion

In this paper, the problem of estimating the fundamental frequency as well as the direction-of-arrival of a desired, periodic signal has been considered, and some new methods based on spatio-temporal filtering have been proposed. The methods are based on optimal filter designs that leave periodic signals of a certain fundamental frequency from a certain direction-of-arrival unchanged while everything else is attenuated as much as possible. The resulting filters are adaptive if the statistics of the observed signal is estimated adaptively, and several incarnations of the ideas have been presented, including single filter and filterbank designs, simplifications based on the assumption that the observed noise signal is white and the filters being infinitely long. The application of the recently introduced iterative adaptive approach to estimation of the involved covariance matrix has also been proposed and investigated. This approach is capable of overcoming the usual limitations on the filter length relative to the number of samples available. That is, with this approach we can estimate the pitch and DOA using fewer samples which is preferable when processing nonstationary signals such as speech. In simulations, the proposed methods outperform state-of-the-art methods under adverse conditions, including the recently proposed maximum likelihood approach which is optimal for white, Gaussian noise and a single periodic signal. More specifically, the spatio-temporal filtering methods outperform the competing methods when multiple periodic signals are present at the same time, something that frequently happens in practice, cf. the well-known cocktail party problem. Finally, experiments on real data in form of a trumpet signal show the applicability of the proposed optimal filtering methods even in scenarios with slight reverberation.

References

[1] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[2] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[3] V. C. Shields, Jr., "Separation of added speech signals by digital comb filtering," S.M. thesis, Mass. Inst. of Technol., Cambridge, MA, USA, 1970.

[4] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 5, pp. 1124–1138, Oct. 1986.

[5] K. Nishi and S. Ando, "An optimal comb filter for time-varying harmonics extraction," *IEICE Trans. Fundamentals*, vol. E81-A, no. 8, pp. 1622–1627, Aug. 1998.

[6] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "On optimal filter designs for fundamental frequency estimation," *IEEE Signal Process. Lett.*, vol. 15, pp. 745–748, 2008.

[7] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 2002.

[8] M. Brandstein and D. Ward, *Microphone Arrays - Signal Processing Techniques and Applications*. New York, NY, USA: Springer-Verlag, 2001.

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

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

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

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

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

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

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

[16] M. Képesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," in *Proc. Hands-Free Speech Commun. Microphone Arrays (HSCMA '08)*, May 2008, pp. 85–88.

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

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

[19] Z. Zhou, M. G. Christensen, J. R. Jensen, and H. C. So, "Joint DOA and fundamental frequency estimation based on relaxed iterative adaptive approach and optimal filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 6812–6816.

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

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

[22] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer, "Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 425–443, Jan. 2010.

[23] W. Roberts, P. Stoica, J. Li, T. Yardibi, and F. A. Sadjadi, "Iterative adaptive approaches to MIMO radar imaging," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 1, pp. 5–20, Feb. 2010.
[24] G.-O. Glentis and A. Jakobsson, "Efficient implementation of iterative adaptive approach spectral estimation techniques," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4154–4167, Sep. 2011.
[25] G.-O. Glentis and A. Jakobsson, "Superfast approximative implementation of the IAA spectral estimate," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 472–478, Jan. 2012.
[26] 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, pp. 3900–3904.
[27] P. Stoica, A. Jakobsson, and J. Li, "Matched-filter bank interpretation of some spectral estimators," *Elsevier Signal Process.*, vol. 66, no. 1, pp. 45–59, 1998.
[28] J. Li and P. Stoica, "An adaptive filtering approach to spectral estimation and SAR imaging," *IEEE Trans. Signal Process.*, vol. 44, no. 6, pp. 1469–1484, Jun. 1996.
[29] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.
[30] 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. Berlin, Germany: Springer-Verlag, 2001, ch. ch. 8, pp. 157–180.
[31] E. A. P. Habets, "Room impulse response generator," Technische Univ. Eindhoven, Tech. Rep., ver. 2.0.20100920, 2010. [Online]. Available: http://home.tiscali.nl/ehabets/rir_generator.html

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, and 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 the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, and at the Friedrich-Alexander Universität Erlangen-Nürnberg in Erlangen, Germany. He has published more than 30 papers in peer-reviewed conference proceedings and journals. Among others, his research interests are digital signal processing, microphone array signal processing, and joint audio-visual signal processing with application to, e.g., enhancement, separation, localization, and tracking of speech and audio sources. In particular, he is interested in parametric analysis, modeling and extraction of such signals. Dr. Jensen has received an individual postdoc grant from the Danish Independent Research Council as well as several travel grants from private foundations.

Mads Græsbøll Christensen (S'00–M'05–SM'11) 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 in the Department of Architecture, Design and Media Technology as Professor in Audio Processing and is head and founder of the Audio Analysis Lab. He was formerly with the Department of Electronic Systems at AAU and has been held visiting positions 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 2 research monographs. His research interests include signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding. Prof. Christensen has received several awards, including the Spar Nord Foundation's Research Prize, a Danish Independent Research Council Young Researcher's Award, and the Statoil Prize, as well as grants from the Danish Independent Research Council and the Villum Foundation's Young Investigator Programme. He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and has previously served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.

Jacob Benesty was born in 1963. He received a Master degree in microwaves from Pierre and Marie Curie University, France, in 1987, and a Ph.D. degree in control and signal processing from Orsay University, France, in April 1991. During his Ph.D. (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecomunications (CNET), Paris, France. From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ, USA. In May 2003, he joined the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, as a Professor. He is also a Visiting Professor at the Technion, in Haifa, Israel, and an Adjunct Professor at Aalborg University, in Denmark and at Northwestern Polytechnical University, in Xi'an, Shaanxi, China. His research interests are in signal processing, acoustic signal processing, and multimedia communications. He is the inventor of many important technologies. In particular, he was the lead researcher at Bell Labs who conceived and designed the world-first real-time hands-free full-duplex stereophonic teleconferencing system. Also, he conceived and designed the world-first PC-based multi-party hands-free full-duplex stereo conferencing system over IP networks. He was the co-chair of the 1999 International Workshop on Acoustic Echo, and Noise Control, and the general co-chair of the 2009 IEEE Workshop on Applications of Signal Processing to Audio, and Acoustics. He is the recipient, with Morgan, and Sondhi, of the IEEE Signal Processing Society 2001 Best Paper Award. He is the recipient, with Chen, Huang, and Doclo, of the IEEE Signal Processing Society 2008 Best Paper Award. He is also the co-author of a paper for which Huang received the IEEE Signal Processing Society 2002 Young Author Best Paper Award. In 2010, he received the "Gheorghe Cartianu Award" from the Romanian Academy. In 2011, he received the Best Paper Award from the IEEE WASPAA for a paper that he co-authored with Chen.

Søren Holdt Jensen (S'87–M'88–SM'00) 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 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, Elsevier Signal Processing and EURASIP Journal on Advances in Signal Processing, and is currently Associate Editor for the IEEE/ACM 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.