Sparse Localization of Harmonic Audio Sources

Adalbjörnsson, Stefan Ingi; Kronvall, Ted; Burgess, Simon; Åström, Karl; Jakobsson, Andreas

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI:
10.1109/TASLP.2015.2497798

2016

LUND UNIVERSITY

Sparse Localization of Harmonic Audio Sources

STEFAN I. ADALBJÖRNSSON, TED KRONVALL, SIMON BURGESS, KALLE ÅSTRÖM, AND A. JAKOBSSON

Published in: IEEE Transactions on Signal Processing
doi:10.1109/TASLP.2015.2497798

Lund 2016

Mathematical Statistics
Centre for Mathematical Sciences
Lund University

Sparse Localization of Harmonic Audio Sources

Stefan I. Adalbjörnsson, *Student member, IEEE*, Ted Kronvall*, *Student member, IEEE*, Simon Burgess, Kalle Åström, *Senior member, IEEE*, Andreas Jakobsson, *Senior member, IEEE*

***Abstract*—In this paper, we propose a novel method for estimating the locations of near- and/or far-field harmonic audio sources impinging on an arbitrary, but calibrated, sensor array. Using a joint pitch and location estimation formed in two steps, we first estimate the fundamental frequencies and complex amplitudes under a sinusoidal model assumption, whereafter the location of each source is found by utilizing both the difference in phase and the relative attenuation of the magnitude estimates. As audio recordings often consist of multi-pitch signals exhibiting some degree of reverberation, where both the number of pitches and the source locations are unknown, we propose to use sparse heuristics to avoid the necessity of detailed a priori assumptions on the spectral and spatial model orders. The method's performance is evaluated using both simulated and measured audio data, with the former showing that the proposed method achieves near-optimal performance, whereas the latter confirms the method's feasibility when used with real recordings.**

***Index Terms*—Multi-pitch estimation, near- and far-field localization, TDOA, block sparsity, convex optimization, ADMM, non-convex sparsity.**

I. Introduction

SOUND localization has been a topic of interest in a wide range of applications for centuries, and is well known to be a difficult problem, especially in a reverberating room environment (see, e.g., [2]–[9], and the references therein). Typically, a source is located in relation to an array of sensors by exploiting the time delay between sensors for when they receive its emitted signal. In the literature, this is referred to as either time of arrival (TOA) estimation, if the time of signal emission is known, or otherwise time difference of arrival (TDOA) estimation, where only the relative time delays are used. Common techniques for delay estimation include different variations on cross-correlation or canonical correlation analysis (CCA), which then allows the sources to be located in a second step using tri- and multi-lateration (see, e.g., [10]) Such estimates may also be further improved by matching the relative received signal gains to a model for signal attenuation. If the source is far from the sensor array, i.e., in the far-field, its range may not be determined due to the lack of curvature of the impinging sound pressure wavefront, which is then approximately planar, making the range estimation problem ill-posed. The scope is then restricted to determining the direction of arrival (DOA) of the source relative to the sensor array for the 2-D case, or determining azimuth and elevation angles for a 3-D scenario. Historically, such methods are not restricted to sound, but are commonly used, in e.g., military applications, with electromagnetic signals (see, e.g., [11]–[13]). Perhaps, partly due to differences in application for near-field and far-field techniques, these problems are often treated separately. In this work, and for our purposes with audio signals, the two problems may indifferently be treated together. A common issue with correlation-based techniques is that of reverberation. Although often described in a temporal sense as a filter for each sensor through which the signal is convoluted [14], it may also be analyzed using a spatial formulation. In principle, reverberation occurs when the original source signal is received together with a number of reflections of it, which are both time delayed and dislocated in space with respect to the original. Localization in reverberant environments is still very much an open topic, although several correlation-based approaches exist which shows some degree of robustness (see, e.g., [4]). By assuming a temporal and spectral parametric structure on the received signals, localization may be improved by jointly forming estimates of location together with the parameters of such structures. This is quite common for audio signals such as voiced speech [14], and many forms of harmonic audio sources, such as stringed, wind, and pitched percussion instruments [15], which typically have lots of structure. At a glance, the spectral distribution of energy for such signals is typically broadband, but further analysis shows that it is in fact dominantly multi-narrowband, and may be well described using the harmonic model, i.e., as a sum of harmonically related sinusoids [16]. Under this assumption, a source's difference in delay and attenuation when received at the different sensors translates into phase shifted and magnitude scaled versions of the original signal. Exploiting this, joint estimation of the DOA and the pitch frequency has been addressed, such as in [17]–[19], wherein the authors consider the estimation of the DOA of a single harmonic sound source using a uniform linear array (ULA) of receiver sensor, typically assuming oracle knowledge of the number of harmonic signals in the sound source. Here, we extend on these works, albeit with some generalizations. We are allowing for an unknown number of near- or far-field harmonic sources, each having an unknown number of harmonics, to impinge on an arbitrary, but calibrated, sensor array, in the presence of some degree of reverberation. This feat is attempted through the use of a sparse recovery framework, which avoids making explicit assumptions on the number of harmonic signals, i.e., the number of pitches, as well as for the number of source locations for each pitch. Instead, only an implicit constraint which controls a lower threshold for acceptable source power

This work was supported in part by the Swedish Research Council, Carl Trygger's foundation, and the Royal Physiographic Society in Lund. This work has been presented in part at the ICASSP 2014 conference [1].

*Corresponding author. Centre for Mathematical Sciences, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden, emails: {sia, ted, simonb, kalle, aj}@maths.lth.se, phone: +4646-222 00 00, fax:+4646-222 42 13.

All authors are affiliated with the Centre for Mathematical Sciences, Lund University

Upon acceptance, all code and data will be made publicly available.

is needed, which may typically be set using some simple heuristics. Sparse recovery frameworks have in earlier works been found to allow high quality estimates for sinusoidal signals; typical examples include [20]–[23], wherein the sparse signal reconstruction from noisy observations were accomplished with the by now well-known sparse least squares (LS) technique. More recently, the technique has been extended to the case of harmonically related audio signals [24], [25]. Using the techniques introduced there, we propose a two-step procedure, first creating a dictionary of candidate pitches to model the harmonic components of the sources, without taking the locations of the sources into account, and then, in a second step, a dictionary of possible locations, including simultaneously near- and far-field locations, to model the observed phase differences, as well as the relative attenuations, of the magnitudes of each sinusoidal component. In terms of computational complexity, the estimation problem in each of the two steps is convex, which thus guarantees convergence, and may be solved using a second order cone (SOC) program. As this is typically quite costly, we introduce a computationally efficient implementation based on the alternating direction method of multipliers (ADMM), which makes the proposed method very managable in an off-line estimation procedure. The remainder of this paper is organized as follows: in the next section, we present the assumed signal model and discuss the imposed restrictions on the sensor array. Then, in section III, we present the proposed pitch and localization estimator. Section IV accounts for the ADMM-based implementation, followed in section V with an evaluation of the presented technique using both simulated and measured audio signals. Finally, we conclude on our work in section VI.

II. Signal model

In this work, we restrict our attention to the localization of complex-valued[1] harmonically related audio signals, consisting of $\tilde{K}$ distinct sources, $x_k(t)$, for $k = 1, \ldots, \tilde{K}$. Each source is thus assumed to consist of L_k harmonically related sinusoids, such that it may be detailed as (see also [16])

$$x_k(t) = \sum_{\ell=1}^{L_k} a_{k,\ell} e^{j\omega_k \ell t} \tag{1}$$

where $\omega_k = 2\pi f_k / f_s$ is the normalized fundamental frequency, with sampling frequency f_s, and with $a_{k,l}$ denoting the complex amplitude of the ℓ:th harmonic.

A. Multi-sensor characteristics in near-field environments

When a source signal impinges on a sensor array, it is both delayed and attenuated, such that at sensor m it may be expressed as

$$x_{k,m}(t) \triangleq \frac{d_{k,1}}{d_{k,m}} x_k(t - \tau_{k,m}) \tag{2}$$

[1]Clearly, the measured audio sources will be real-valued, but to simplify notation and in order to reduce complexity, we will here initially compute the discrete-time analytic signal versions of the measured signals, whereafter all processing is done on these signals (see also [16], [26]).

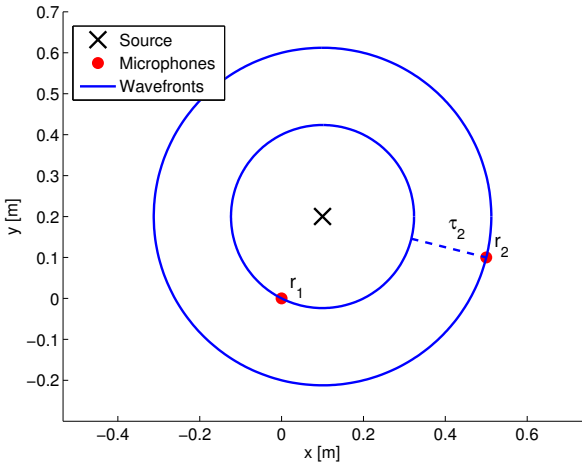

Fig. 1. Illustration of a two sensor scenario, with spherical wavefronts propagating from the source. The dashed line shows the scaled TDOA of the second sensor with respect to the first sensor, i.e., τ_2.

where $d_{k,m}$ denotes the sensor-source distance, i.e.,

$$d_{k,m} = \|\mathbf{s}_k - \mathbf{r}_m\|_2 \tag{3}$$

with $\mathbf{s}_k$ and $\mathbf{r}_m$ denoting the location coordinates of the k:th source and the m:th sensor, respectively, and $\|\cdot\|_2$ the Euclidean norm. Thus, (2) accounts for the approximative attenuation of the signal when propogating in space, according to the free-space path loss model. Furthermore, $\tau_{k,m}$ denotes the propagation delay, i.e., the TDOA, relative to a selected reference sensor, say $m = 1$, so that

$$\tau_{k,m} = c^{-1}\left(d_{k,m} - d_{k,1}\right) \tag{4}$$

for $m = 1, \ldots, M$, where $\tau_{k,1} \triangleq 0$, with c denoting the propagation velocity. An illustration of this is shown in Figure 1, for the case of a single source and two sensors. When recording audio, we often obtain multi-pitch signals of the type

$$x(t) = \sum_{k=1}^{\tilde{K}} x_k(t) \tag{5}$$

which may be either a single source in the physical environment emitting multiple pitch signals, such as an instrument playing a chord, or multiple sources in the physical environment each emitting a single pitch, such as multiple speakers talking at the same time from different locations. We may also receive a combination of these two types. Without loss of generality, we will hereafter term a source as a spatio-temporal object which has a unique combination of fundamental frequency and location. Two sources may thus have the same fundamental frequency or the same location in space, although not both. This has rather large implications when considering reverberation, where we, apart from the original source, also receive a large number of reflections of it, each reflection having highly similar spectral content, albeit differently attenuated and delayed, i.e., having different magnitudes and phases. All reflections will thus be modeled

as separate sources, which implies that under such a model assumption $\tilde{K}$ generally becomes very large. If not seen as separate sources, however, the localization of the original source will become biased by the interference caused from its reflections. To see this, consider for example a sinusoid with frequency ω, magnitude a_1, and phase φ_1, measured in superimposition with its $S-1$ reverberating reflections, having magnitudes $a_2, \ldots, a_S$, and phases $\varphi_2, \ldots, \varphi_S$. For the mth sensor, the measured (noise-free) signal becomes

$$x_m(t) = \sum_{s=1}^{S} a_s e^{-j(\omega t + \varphi_s)} \triangleq b e^{-j(\omega_0 t + \psi)} \tag{6}$$

i.e., a single sinusoid with magnitude $b \in \mathbb{R}_+$ and phase $\psi \in [-\pi, \pi)$, generally being different from the original source. Thus, if trying to estimate the TDOA using phase estimates without taking all reflections into account, for instance by using a correlation-based measure, then only the biased phase, ψ, would be obtained. However, separation of all reflections for all fundamental frequencies is a quite difficult problem, and in this work, we propose to split the estimation procedure into two subproblems. In the first, we find the present fundamental frequencies, and then for each of these we separate the original source(s) from its reflections. To that end, consider $K \leq \tilde{K}$ as the number of unique fundamentals. The noisy signal measured at sensor m may thus be expressed as

$$y_m(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} b_{k,\ell,m} e^{j\omega_k \ell t} + e_m(t) \tag{7}$$

where the TDOA and attenuation of all S_k reflections of the k:th pitch, for overtone ℓ and sensor m, is gathered in the complex amplitude of the signal, $b_{k,\ell,m}$ using (2) in the same manner as in (6), i.e.,

$$b_{k,\ell,m} = \sum_{s=1}^{S_k} a_{k,\ell,s} \frac{d_{k,1,s}}{d_{k,m,s}} e^{-j\omega_k \ell \tau_{k,m,s}} \tag{8}$$

where $a_{k,\ell,s}, d_{k,m,s}$, and $\tau_{k,m,s}$ denote the amplitude, the distance to the mth sensor, and the TDOA for the sth reflection, respectively. Thus, as $\tilde{K} = \sum_{k=1}^{K} S_k$, the estimation procedure first finds the K active fundamentals, whereafter for each one, the original source is separated from its reflections. This approach offers great simplification in contrast to decoupling all $\tilde{K}$ sources simultaneously. To simplify presentation, and without loss of generality, we will here restrict our attention to the case when all sources and signals are restricted to a 2-D plane, i.e., $\mathbf{s} \in \mathbb{R}^2$ and $\mathbf{r} \in \mathbb{R}^2$.

B. Avoiding spatial aliasing in arbitrary array geometries

In the literature, keeping below half wavelength sensor spacing is generally preferred to avoid spatial aliasing, although some methods of circumventing this have been published, see e.g. [27]. In this work, we assume a calibrated, although arbitrary, sensor array, without requiring it to satisfy the pairwise half wavelength spacing. We will therefore briefly examine the spatial aliasing effect in the near-field environment, which is the phase difference ambiguity between sensors,

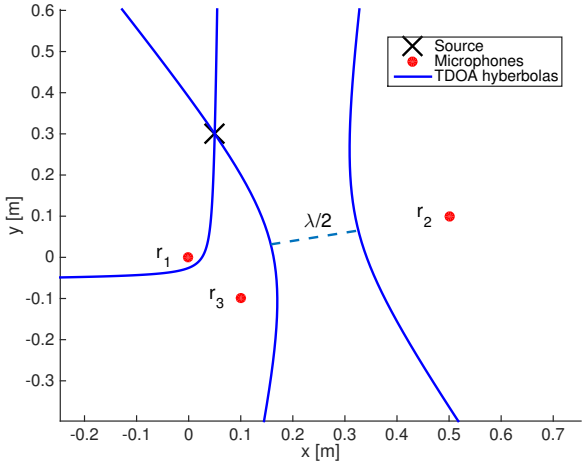

Fig. 2. TDOA hyperbolas representing all feasible locations of a single source received by three sensors. As $\|\mathbf{r}_2 - \mathbf{r}_1\| > \lambda/2$, spatial aliasing yields another hyperbola of feasible locations. And yet, in this case, there exists only one intersection between the hyperbolas and so the estimate may still be obtained unambiguously.

resulting when the solution may map to several feasible source locations. To that end, consider a reverberation-free, delayed, and attenuated complex amplitude from a single sinusoidal signal, b. Naturally,

$$b_m = \frac{d_1}{d_m} a e^{-j\omega\tau_m} = \frac{d_1}{d_m} a e^{-j(\omega\tau_m + k2\pi)} \tag{9}$$

and thus the mapping between phase and TDOA is ambiguous for any $k \in \mathbb{Z}$. Considering a given TDOA, and by combining (3) and (4), one will note that any source $\mathbf{s}$ located on a half-space of an hyperbolic curve, i.e.,

$$\tau_m c = \|\mathbf{s} - \mathbf{r}_m\|_2 - \|\mathbf{s} - \mathbf{r}_1\|_2 \tag{10}$$

is a feasible location. To obtain a unique solution, we add additional sensors, and we may thus form new sensor pairs yielding new hyperbolas, where the feasible solution set will be restricted by the intersection of these curves. Ambiguity may arise when, for each sensor pair, there exist another TDOA (and thus another k) which fulfills (9), giving rise to an additional hyperbolic curve of feasible points, also intersecting the hyperbolas for other sensor pairs. To identify such ambiguous cases, we first show that a feasible TDOA is restricted to an interval. Using the triangle inequality,

$$|\tau_m c| = \left| \|\mathbf{s} - \mathbf{r}_m\|_2 - \|\mathbf{s} - \mathbf{r}_1\|_2 \right| \leq \|\mathbf{r}_m - \mathbf{r}_1\|_2 \tag{11}$$

it is directly implied that the TDOA must satisfy

$$\tau_m c \in \left[-\|\mathbf{r}_m - \mathbf{r}_1\|_2 , \|\mathbf{r}_m - \mathbf{r}_1\|_2 \right] \tag{12}$$

i.e., is restricted by the sensor-sensor distance. And so, using (9), an estimate of $\arg b \in [-\pi, \pi]$ will map to any TDOA

$$\tau_m c = \frac{\lambda \arg b}{2\pi} + \lambda k \in \left[-\|\mathbf{r}_m - \mathbf{r}_1\|_2 , \|\mathbf{r}_m - \mathbf{r}_1\|_2 \right] \tag{13}$$

where $k \in \mathbb{Z}$, and $\lambda = 2\pi c/\omega$ is the wavelength of the signal. Therefore, if the sensors are spaced by less than

$\lambda/2$, the feasible τ_m is unique, and there is no ambiguity in the resulting estimates. If instead some sensors are spaced further apart than $\lambda/2$, then, for all such sensor pairs, there will be more than one feasible TDOA, thereby yielding as many hyperbolas indicating feasible source locations, with a minimum distance of $\lambda/2$ apart. Our main argument to relax the halv wavelength spacing limit is that, when using sufficiently many sensors, the feasible source locations are restricted to the intersection of many hyperbolas, which will, with a high probability, yield a unique solution. Consider an example illustrated in Figure 2, where a single source emits a 1000 Hz signal, which is recorded by three sensors. As shown in the figure, between sensors one and three, which are less than $\lambda/2$ apart, the source gives a single TDOA and a corresponding hyperbola, where the source may be located. Between sensors one and two, which are spaced by more than $\lambda/2$ apart, a second TDOA is feasible, λ/c apart from the true one, also fulfilling (13). However, as shown in the figure, the combined hyperbolas coincide in only a single feasible location, thus still allowing for an unambiguous estimate of the source location. Furthermore, for pitch signals, each overtone will yield a separate set of hyperbolas, which all must intersect to the same location, which further helps to avoid ambiguity. Modeling the attenuation between sensors also helps to avoid ambiguity. Examining the magnitude of the the complex amplitude in (9), we find that

$$|b_m| = \frac{d_1}{d_m}|a| \tag{14}$$

for each pair, consisting of the first and the m:th microphone, which limits $\mathbf{s}$ to lie on a circle. Using the same arguments as above, a feasible source location in terms of attenuation is thus the intersection of circles for all microphone pairs, and will further contribute to avoid spatial aliasing. Even if, despite of intersecting the feasible solutions for all harmonics in terms of both delay and attenuation, ambiguities still remain, then as more sensors are added to the array the set of possible locations quickly becomes small, and a unique solution generally exists, even if not guaranteed. We thus deem that the imposed restriction on the array's geometry is mild.

III. Joint Pitch and Location Estimation

We proceed to detail the proposed two-step procedure to form reliable estimates of both the pitches and locations of the sources impinging on the array, without assuming detailed model knowledge of either the number of sources, K, the number of overtones for each source, L_k, the number of reflections experienced due to a possibly reverberant environment, S_k, or requiring knowledge about whether sources are far- or near-field. In the first step, the magnitudes, phases, fundamental frequencies, and model orders of the present pitches are estimated, and subsequently, in the second step, the phase estimates are used to find the locations of these sources. Let

$$\mathbf{\Phi} = \left\{ \{b_{k,\ell,m}\}_{\substack{\ell=1,\ldots,L_k \\ m=1,\ldots,M}}, \omega_k, L_k \right\}_{k=1,\ldots,K} \tag{15}$$

denote the set of unknown parameters to be determined in the first step. Minimizing the squared model residual in (7), an estimate of $\mathbf{\Phi}$ may thus be formed as

$$\hat{\mathbf{\Phi}} = \arg\min_{\mathbf{\Phi}} \sum_{t=1}^{N} \sum_{m=1}^{M} \left| y_m(t) - \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} b_{k,\ell,m} e^{j\omega_k \ell t} \right|^2 \tag{16}$$

Clearly, given the dimensionality of the problem, and the required model order estimation steps in order to determine K and L_k, this is a non-trivial problem, and needs to be modified to allow for an efficient solution, as is detailed below. Moving over to the second step, the found magnitude and phase estimates, $\hat{b}_{k,l,m}$, are exploited to form estimates of the source locations. To that end, let

$$\mathbf{\Psi}_k = \left\{ \{a_{k,\ell,s}\}_{\ell=1,\ldots,L_k}, \mathbf{s}_s \right\}_{s=1,\ldots,S_k} \tag{17}$$

be the amplitudes and coordinates for a present fundamental frequency k. The locations may be determined by minimizing the squared model residual in (8), i.e.,

$$\hat{\mathbf{\Psi}}_k = \arg\min_{\mathbf{\Psi}_k} \sum_{\ell=1}^{\hat{L}_k} \sum_{m=1}^{M} \left| \hat{b}_{k,\ell,m} - \sum_{s=1}^{S_k} a_{k,\ell,s} d_{k,m,s}^{-1} e^{-j\omega_k \ell \tau_{k,m,s}} \right|^2 \tag{18}$$

where $\tau_{k,m,s}$ and $d_{k,m,s}$ are functions of the location $\mathbf{s}_s$, as defined in (3) and (4). As before, this minimization is also non-trivial, requiring an estimate of S_k, and also needs to be modified to allow for a reasonably efficient solution. In the following, we will elaborate on the proposed modifications of the above minimizations. In order to do so, we first extend the sparse pitch estimation algorithm presented in [24], [25] to allow for multiple measurement vectors. In the second minimization, we then introduce a similar sparsity pattern to solve the localization problem. We begin by examining the extended pitch estimation algorithm.

A. Step 1: Sparse pitch estimation

Define the measurement matrix

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}(1) & \ldots & \mathbf{y}(N) \end{bmatrix}^T \tag{19}$$

where

$$\mathbf{y}(t) = \begin{bmatrix} y_0(t) & \ldots & y_{M-1}(t) \end{bmatrix}^T \tag{20}$$

denotes a sensor snapshot for each time point $t = 1, \ldots, N$, with $(\cdot)^T$ being the transpose. The measurements may then be concisely expressed as

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{W}_k \mathbf{B}_k + \mathbf{E} \tag{21}$$

where $\mathbf{E}$ denotes the combined noise term constructed similar to $\mathbf{Y}$, and

$$\mathbf{W}_k = \begin{bmatrix} \mathbf{w}_k^1 & \ldots & \mathbf{w}_k^{L_k} \end{bmatrix} \tag{22}$$

$$\mathbf{w}_k = \begin{bmatrix} e^{j\omega_k} & \ldots & e^{j\omega_k N} \end{bmatrix}^T \tag{23}$$

$$\mathbf{B}_k = \begin{bmatrix} \mathbf{b}_{k,1} & \ldots & \mathbf{b}_{k,L_k} \end{bmatrix}^T \tag{24}$$

$$\mathbf{b}_{k,\ell} = \begin{bmatrix} b_{k,\ell,1} & \ldots & b_{k,\ell,M} \end{bmatrix}^T \tag{25}$$

Reminiscent to the sparse estimation framework proposed in [20], we form an extended dictionary of feasible fundamental frequencies, $\omega_1, \ldots, \omega_P$, where $P \gg K$, being chosen so large that K of these will coincide reasonably well with the true pitches in the signal. In the same manner, the number of harmonics of each pitch is extended to an arbitrary upper level, say L_{max}, for all dictionary elements. The signal model may thus be expressed as

$$\mathbf{Y} = \sum_{p=1}^{P} \mathbf{W}_p \mathbf{B}_k + \mathbf{E} = \mathcal{W}\mathcal{B} + \mathbf{E} \tag{26}$$

where the block dictionary matrices are formed by stacking the matrices such that

$$\mathcal{W} = \left[\begin{array}{ccc} \mathbf{W}_1 & \ldots & \mathbf{W}_P \end{array} \right] \tag{27}$$

$$\mathcal{B} = \left[\begin{array}{ccc} \mathbf{B}_1^T & \ldots & \mathbf{B}_P^T \end{array} \right]^T \tag{28}$$

Note from (26) that if the element (ℓ, r) of the matrix $\mathbf{B}_k$ is non-zero, the frequency $\ell\omega_k$ is present in the signal at sensor r. Furthermore, since we assume all sensors to receive essentially the same signal, although time-delayed, one may assume that for a harmonic signal, the rows off a non-zero $\mathbf{B}_k$ will either be non-zero, implying that the harmonic ℓ is present in the pitch, or zero, if the harmonic is missing. An appropriate criterion, that promotes a combination of model to data fit and the sparsity pattern just described, may thus be formed as

$$\underset{\mathcal{B}}{\text{minimize}} \left\{ \frac{1}{2} \left\| \mathbf{Y} - \mathcal{W}\mathcal{B} \right\|_{\mathcal{F}}^2 + \lambda \sum_{p=1}^{P} \sum_{\ell=1}^{L_p} \left\| \mathbf{b}_{p,\ell} \right\|_2 + \sum_{p=1}^{P} \gamma_p \left\| \mathbf{B}_p \right\|_{\mathcal{F}} \right\} \tag{29}$$

where two different kinds of group sparsities are imposed, and with $\|\cdot\|_{\mathcal{F}}$ denoting the Frobenius norm. This can be seen to be a generalization of the sparse group lasso to the multiple measurement case (see also [25], [28]). Here, the double sum of 2-norms in the second entry of the minimization should enforce sparsity in the solution in the rows of $\mathcal{B}$, and ideally only have as many non-zero rows as there are sinusoids in the signal. The third entry makes the solution (matrix) block sparse over the candidate pitches, penalizing the number of pitches with non-zero magnitude in the signal, ideally making them as many as there are pitches in the signal, i.e., K. Given an optimal point, $\hat{\mathcal{B}}$, the number of pitches is thus estimated as the number of non-zero matrices $\hat{\mathbf{B}}_k$, and, for each pitch, the number of harmonics, L_k, is estimated as the number of non-zero rows. The user parameters $\lambda, \gamma_p \in \mathbb{R}_+$ weighs the fit of the solution to its vector and matrix sparsity, respectively. It is well known (see, e.g., [29]) that the amplitudes in the sparse estimate will be increasingly biased towards zero as sparse regularizers are increased. As we here intend to use both the estimated phases and the magnitudes, we propose to refine the amplitude estimates using a reweighting scheme similar to the one presented in [30]. This is accomplished by iteratively solving (29), such that at iteration $j+1$, one updates

$$\gamma_p^{(j+1)} = \frac{\gamma_p^{(0)}}{\left\| \hat{\mathbf{B}}_p^{(j)} \right\|_{\mathcal{F}} + \varepsilon} \tag{30}$$

where $\hat{\mathbf{B}}_p^{(j)}$ is block p of the optimal point for iteration j, and all $\gamma_p^{(0)}$ are set to be equal in the first iteration. As a result, the block matrices, $\hat{\mathbf{B}}_p^{(j)}$, which have a small Frobenius norm at iteration j will be penalized harder in the next step, whereas the ones that have a larger Frobenius norm will be penalized less, and as a result reducing the bias. The resulting algorithm can be seen as a sequence of iterative convex programs to approximate the concave $\log(\sum_{p=1}^{P} \gamma_p^{(0)} \left\| \mathbf{B}_p \right\|_{\mathcal{F}} + \varepsilon)$ penalty function [31], where ϵ is chosen as a small number to avoid numerical difficulties. The introduction of the reweighting yields sparser estimates due to the introduction of the log penalty [30], [32], and the resulting technique may be viewed as an alternative to using an information criterion (as was done in [25], to avoid spurious peaks caused by the signal model and data miss-match).

It is worth noting that as we are here focusing on localization, we have selected to use a somewhat simplistic audio model that ignores several important features in harmonic audio signals, such as issues of inharmonicities, pitch halvings and doublings, and of the commonly occurring forms of amplitude modulation exhibited by most audio sources (see also [16]). Clearly, the used model could be refined reminiscent to models such as the one used in [25], [33], introducing a total variation penalty to each column of $\mathcal{B}$, and/or using an uncertainty volume to allow for inharmonicity. However, for localization purposes, these issues are of less concern, as halvings/doublings and/or amplitude modulations will not affect the below localization procedure more than marginally. Inharmonicity is more pressing, but we have in our numerical studies found that given the size of the calibration errors, the inharmonicity is not affecting the solution significantly, and in the interest of reducing the complexity, we have here opted to exclude this aspect from the estimator.

As for the selection of the tuning parameters, one may use, for example, cross validation techniques, although it may be noted that, in high SNR cases, one can often get good results by simply inspecting the periodogram and by then setting the tuning parameters appropriately (see also [25] for a further discussion on this issue). Furthermore, we note that in the case of different noise variances at each sensor in the array, the Frobenius norm in the first entry of the minimization criterion may be replaced with a weighed Frobenius norm. Finally, we note that non-Gaussian noise distributions can also be used as long as the negative log-likelihood is convex.

B. Step 2: Sparse localization

According to the signal model (7), $\hat{\mathcal{B}}$ will inherently contain the TDOA and attenuation for all reflections of any fundamental frequency present in the signal, which enables a range of post-processing steps to, for instance, estimate position, track, and/or calibrate the sensors. Here, we limit our attention to estimating the source positions. Let $\hat{\mathcal{B}}$ denote the solution

obtained from minimizing (29), and consider a scenario where the sources are well separated in their pitch frequencies, and, initially, suffering from negligible reverberation, implying that $S_1 = \ldots = S_P = 1$. Then, the minimization in (18) may be seen as a generalization of the time-varying amplitude modulation problem examined in [34] (see also [13]) to the case of several realizations of the same signal, sampled at irregular time points, and with a different initial phase for each realization. Reminiscent to the solution presented in [13, p. 186], one may thus find the source locations, for far-field signals, for every pitch p with non-zero amplitudes in $\mathbf{B}_p$, as

$$\hat{\mathbf{s}}_p = \arg\max_{\mathbf{s}_p} \sum_{\ell=1}^{L_p} \left| \sum_{m=1}^{M} \hat{b}_{p,\ell,m}^2 e^{-j2\omega_p \ell \tau_{p,\ell,m}} \right|^2 \tag{31}$$

where the TDOAs $\tau_{p,\ell,m}$ are found as a function of the source location $\mathbf{s}_p$, using (4). This minimization may be well approximated by 1-D searches over range and DOA (or over range, azimuth, and elevation in the 3-D case). Considering also reverberating room environments, wherein each of the pitches may appear as originating from many different locations, the minimization needs to be extended to allow for varying number of reflections, S_k. To allow for such reflections, we proceed to model every non-zero amplitude block from the pitch estimation step as

$$\mathbf{B}_k = \sum_{s=1}^{S_k} \operatorname{diag}(\mathbf{a}_{k,s}) \, \mathbf{U}_{k,s} + \boldsymbol{\mathcal{E}}_k \tag{32}$$

with $\operatorname{diag}(\mathbf{x})$ denoting a diagonal matrix with the vector $\mathbf{x}$ along its diagonal, $\boldsymbol{\mathcal{E}}_k$ the combined noise term constructed in the same manner as $\mathbf{B}_k$, and

$$\mathbf{U}_{k,s} = \begin{bmatrix} \mathbf{u}_{k,s}^1 & \ldots & \mathbf{u}_{k,s}^{\hat{L}_k} \end{bmatrix} \tag{33}$$

$$\mathbf{u}_{k,s} = \begin{bmatrix} \frac{e^{j\omega_k \tau_{k,1,s}}}{1} & \ldots & \frac{e^{j\omega_k \tau_{k,M,s}}}{d_{k,M,s}/d_{k,m,s}} \end{bmatrix}^T \tag{34}$$

$$\mathbf{a}_{k,s} = \begin{bmatrix} \mathbf{a}_{k,1,s} & \ldots & \mathbf{a}_{k,\hat{L}_k,s} \end{bmatrix}^T \tag{35}$$

where $\tau_{k,m,s}$ and $d_{k,m,s}$ are related to the source location as given by (3) and (4), respecively. Analogously to the above procedure for the pitch estimation, we then extend the dictionary of feasible source locations for the kth source, $\mathbf{s}_1, \ldots, \mathbf{s}_{S_k}$, onto a grid of $Q \gg S_k$ candidate locations $\mathbf{s}_q$, for $q = 1, \ldots, Q$, with Q chosen large enough to allow some of the introduced dictionary elements to coincide, or closely so, with the true source locations in the signal. Clearly, this may force Q to be very large. Striving to keep the size of the dictionary as small as possible, we consider grid points in polar coordinates, with equal resolution for all considered DOAs, and linearly spaced grid points over the distance in each DOA. Thus, we get a denser grid in the close proximity to the sensor array, where the resolution capacity is highest, and then a less and less dense grid for sources further away from the array. Finally, to also allow for far-field sources, one may include one dictionary element for each direction at an infinite range, for which, naturally, the attenuation effect may be disregarded, i.e., $d_{k,m,s} \triangleq 1$ for all sensors. Thus, we may estimate the source locations for the k:th pitch using a sparse modelling framework as

$$\underset{\mathbf{a}_{k,1},\ldots,\mathbf{a}_{k,Q}}{\text{minimize}} \left\{ \frac{1}{2} \left\| \mathbf{B}_k - \sum_{q=1}^{Q} \operatorname{diag} \mathbf{a}_{k,q} \mathbf{U}_{k,q} \right\|_{\mathcal{F}}^2 + \sum_{q=1}^{Q} \kappa_q \left\| \mathbf{a}_{k,q} \right\|_2 + \rho \sum_{q=1}^{Q} \left\| \mathbf{a}_{k,q} \right\|_1 \right\} \tag{36}$$

where, again, two types of sparsity is imposed on the solution. The 2-norm penalty term imposes sparsity to the blocks $\mathbf{a}_{k,q}$, i.e., penalizing the number of source locations present in the signal. Furthermore, the 1-norm term penalizes the number of harmonics, to allow for cases when some sources may have missing harmonics. Thus, here the number of sources is estimated as the number of nonzero blocks in an optimal point and any zero elements within a block corresponding to a missing harmonic. Here, $\kappa_q, \rho \in \mathbb{R}_+$ are tuning parameters, controlling the amount of sparsity and the weight between sparsity in pitches and in harmonics, respectively, whereas the factor ρ is only used if two sources share the same fundamental frequency but differ in which harmonics are present. Finally, κ_q may be updated in the same manner as described in section III.A. As shown in the following section, the optimization problem in (29) and (36) are equivalent, so these tuning parameters may be set in a similar fashion.

IV. ADMM IMPLEMENTATION

It is worth noting that both the minimization in (29) and (36) are convex, as the tuning parameters are non-negative and all the functions are convex. Their solutions may thus be found using standard convex minimization techniques, e.g., using CVX [35], [36], SeDuMi [37], or SDPT3 [38]. Regrettably, such solvers will scale poorly both with increasing data length, the use of a finer grid for the fundamental frequencies, and with the number of sensors. Furthermore, such implementations are unable to utilize the full structure of the minimization, and may, as a result, be computationally cumbersome in practical situations. To alleviate this, we proceed to formulate a novel ADMM re-formulation of the minimizations, offering efficient and fast implementations of both minimizations. For completeness and to introduce our notation, we briefly review the main steps involved in an ADMM (we refer the reader to [39], [40] for further details on the ADMM). Considering the convex optimization problem

$$\underset{\mathbf{z}}{\text{minimize}} \; f(\mathbf{z}) + g(\mathbf{z}) \tag{37}$$

where $\mathbf{z} \in \mathbb{R}^p$ is the optimization variable, with $f(\cdot)$ and $g(\cdot)$ being convex functions. Introducing the auxiliary variable, $\mathbf{u}$ (37) may equivalently be expressed as

$$\underset{\mathbf{z},\mathbf{u}}{\text{minimize}} \; f(\mathbf{z}) + g(\mathbf{u}) \quad \text{subject to } \mathbf{z} - \mathbf{u} = \mathbf{0} \tag{38}$$

since at any feasible point $\mathbf{z} = \mathbf{u}$. Under the assumption that there is no duality gap, which is true for the here considered minimizations, one may solve the optimization problem via the dual function defined as the infimum of the augmented Lagrangian, with respect to $\mathbf{x}$ and $\mathbf{z}$, i.e., (see also [39])

$$L_\mu(\mathbf{z}, \mathbf{u}, \mathbf{d}) = f(\mathbf{z}) + g(\mathbf{u}) + \mathbf{d}^T(\mathbf{z} - \mathbf{u}) + \frac{\mu}{2} ||\mathbf{z} - \mathbf{u}||_2^2$$

Algorithm 1 The ADMM algorithm

1: Initiate $\mathbf{z} = \mathbf{z}_0, \mathbf{u} = \mathbf{u}_0$, and $k = 0$
2: **repeat**
3: $\quad \mathbf{z}_{k+1} = \underset{\mathbf{z}}{\operatorname{argmin}}\, f(\mathbf{z}) + \frac{\mu}{2}||\mathbf{z} - \mathbf{u}_k - \mathbf{d}_k||_2^2$
4: $\quad \mathbf{u}_{k+1} = \underset{\mathbf{u}}{\operatorname{argmin}}\, g(\mathbf{u}) + \frac{\mu}{2}||\mathbf{z}_{\mathbf{k+1}} - \mathbf{u} - \mathbf{d}_k||_2^2$
5: $\quad \mathbf{d}_{k+1} = \mathbf{d}_k - (\mathbf{z}_{\mathbf{k+1}} - \mathbf{u}_{\mathbf{k+1}})$
6: $\quad k \leftarrow k + 1$
7: **until** convergence

The ADMM does this by iteratively maximizing the dual function such that at step $k+1$, one minimizes the Lagrangian for one of the variables, while holding the other fixed at its most recent value, i.e.,

$$\mathbf{z}_{k+1} = \arg\min_{\mathbf{z}} L_\mu(\mathbf{z}, \mathbf{u}_k, \mathbf{d}_k) \tag{39}$$

$$\mathbf{u}_{k+1} = \arg\min_{\mathbf{u}} L_\mu(\mathbf{z}_{k+1}, \mathbf{u}_k, \mathbf{d}_k) \tag{40}$$

Finally, one updates the dual variable by taking a gradient ascent step to maximize the dual function, resulting in

$$\tilde{\mathbf{d}}_{k+1} = \tilde{\mathbf{d}}_k - \mu\left(\mathbf{z}_{k+1} - \tilde{\mathbf{d}}_{k+1}\right) \tag{41}$$

where μ is the dual variable step size. The general ADMM steps are summarized in Algorithm 1, using the scaled version of the dual variable $\mathbf{d}_k = \tilde{\mathbf{d}}/\mu$, which is more convenient for implementation. Thus, in cases when steps 3 and 4 of Algorithm 1 may be carried out more efficiently than for the original problem, the ADMM may be useful to form an efficient implementation of the considered minimization.

It may be noted that the minimizations in (29) and (36) are rather similar, both containing an affine function in a Frobenius norm, as well as a sum of the norm of different subset of the variable. In fact, by using the vec operation, i.e., vectorization, both minimizations may be shown to be equivalent with the problem

$$\underset{\mathbf{z}}{\text{minimize}}\left\{\frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 + \gamma\sum_{k=1}^{P}\|\mathbf{z}_k\|_2 + \delta\sum_{k=1}^{P}\sum_{g=1}^{G_k}\|\mathbf{z}_{k,g}\|_2\right\} \tag{42}$$

where the complex variable $\mathbf{z}$ is given as

$$\mathbf{z} = \begin{bmatrix}\mathbf{z}_1^T & \dots & \mathbf{z}_P^T\end{bmatrix}^T \tag{43}$$

$$\mathbf{z}_k = \begin{bmatrix}\mathbf{z}_{k,1}^T & \dots & \mathbf{z}_{k,G_k}^T\end{bmatrix}^T \tag{44}$$

where each $\mathbf{z}_k$ and $\mathbf{z}_{k,g}$ denote complex vectors with G_k and O elements, respectively. For the minimization in (29), this implies that

$$\mathbf{y} = \text{vec}(\mathbf{Y}) \tag{45}$$

$$\mathbf{z} = \text{vec}(\mathcal{B}) \tag{46}$$

$$\mathbf{A} = \mathbf{I} \otimes \mathcal{W} \tag{47}$$

where $\otimes$ and $\mathbf{I}$ denote the Kronecker product and an M-dimensional identity matrix, respectively, with G_k being equal to the number of harmonics, L_k, and O equals the number of sensors, M. Similarly, for the minimization in (36),

$$\mathbf{y} = \text{vec}(\mathbf{B}_p) \tag{48}$$

$$\mathbf{z} = \mathbf{a}_k \tag{49}$$

$$\mathbf{A} = \tilde{\mathbf{V}}_k \tag{50}$$

where

$$\mathbf{a}_k = \begin{bmatrix}\mathbf{a}_{k,1}^T & \dots & \mathbf{a}_{k,Q}^T\end{bmatrix}^T \tag{51}$$

$$\tilde{\mathbf{V}}_k = \begin{bmatrix}\tilde{\mathbf{V}}_{k,1} & \dots & \tilde{\mathbf{V}}_{k,Q}\end{bmatrix} \tag{52}$$

and $\mathbf{V}_{k,q} = \mathbf{U}_{k,q} \otimes \mathbf{I}$, with $\tilde{\mathbf{V}}_{k,q}$ being formed by removing all columns from $\mathbf{V}_{k,q}$ that correspond to zeros in the vector $\text{vec}(\text{diag}(\mathbf{a}_{k,q}))$, and G_k being equal to L_k and O equals 1. Thus, we can formulate an ADMM solution for (42) that solves both problem (29) and (36). To that end, defining

$$f(\mathbf{z}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 \tag{53}$$

$$g(\mathbf{u}) = \gamma\sum_{k=1}^{P}\|\mathbf{u}_k\|_2 + \delta\sum_{k=1}^{P}\sum_{g=1}^{Q_k}\|\mathbf{u}_{k,g}\|_2 \tag{54}$$

yields a quadratic problem in step 3 in Algorithm 1, with a closed form solution given by

$$\mathbf{z}_{k+1} = \left(\mu I + \mathbf{A}^H\mathbf{A}\right)^{-1}\left(\mu\left(\mathbf{u}_k - \mathbf{d}_k\right) + \mathbf{A}^H\mathbf{y}\right)$$

with $(\cdot)^H$ denoting the Hermitian transpose, whereas in step 4, by solving the sub-differential equations (see [25] for further details), one obtains

$$\mathbf{u}_{k+1} = \mathcal{S}^o\left(\mathcal{S}^i\left(\mathbf{z}_k - \mathbf{d}_k, \kappa/\mu\right), \delta/\mu\right) \tag{55}$$

where the shrinkage operators $\mathcal{S}^o$ and $\mathcal{S}^i$ are defined using the vector shrinkage operator $\mathcal{S}$, defined for any vector $\mathbf{v}$ and positive scalar ξ such that

$$\mathcal{S}(\mathbf{v}, \xi) = \mathbf{v}\left(1 - \xi/||\mathbf{v}||_2\right)^+ \tag{56}$$

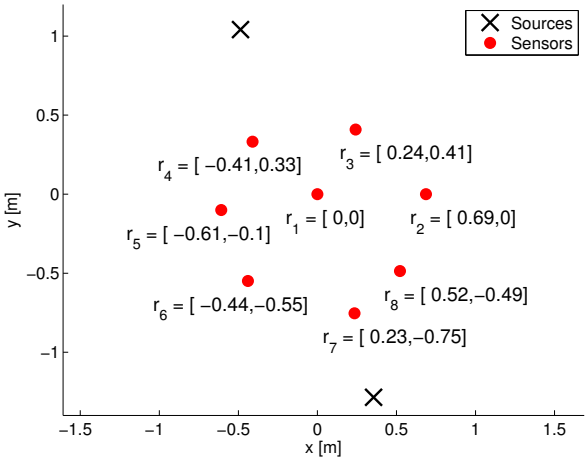

Fig. 3. The two-source and eight-sensor layout in 2-D. The position of each sensor, shown in the plot with carthesian coordinates as $r_m = [x, y]$, was obtained in an *a priori* calibration step.

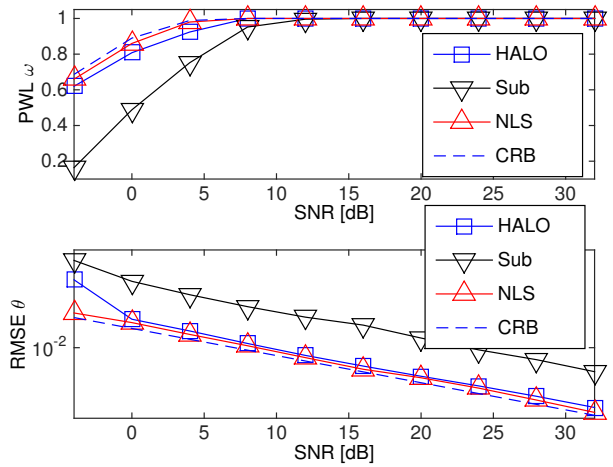

Fig. 4. The PWL and RMSE for a single-pitch signal as compared with the optimal performance of an estimator reaching the CRB.

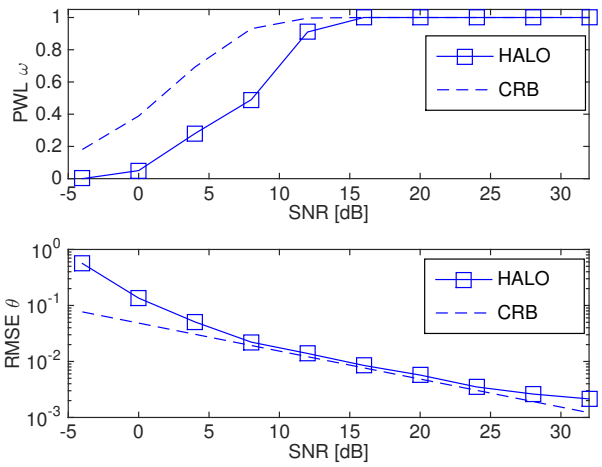

Fig. 5. The PWL and RMSE for a multi-pitch signal with two pitches, as compared to the corresponding CRB.

where $(\cdot)^+$ is the positive part of the scalar, and

$$\mathcal{S}(\mathbf{z},\xi)^o = \left[\begin{array}{ccc} \mathcal{S}^T(\mathbf{z}_1,\xi) & \dots & \mathcal{S}^T(\mathbf{z}_P,\xi) \end{array}\right]^T \tag{57}$$

$$\mathcal{S}(\mathbf{z},\xi)^i = \left[\begin{array}{ccccc} \mathcal{S}^T(\mathbf{z}_{1,1},\xi) & \dots & \mathcal{S}^T(\mathbf{z}_{1,G_1},\xi) & \dots & \\ \mathcal{S}^T(\mathbf{z}_{P,1},\xi) & \dots & \mathcal{S}^T(\mathbf{z}_{P,G_P},\xi) & & \end{array}\right]^T \tag{58}$$

The resulting algorithm is here termed the Harmonic Audio LOcalization using block sparsity (HALO) estimator.

V. Numerical comparisons

We proceed to examine the performance of the proposed estimator using both synthetic and measured audio signals, initially examining the performance using simulated audio signals. In the first examples, we limit ourselves to the case of letting a far-field signal impinge on a ULA. Figure 4 shows the percentage within limits (PWL), defined as the ratio of pitch estimates within a limit of ± 0.1 Hz from the true pitch, and the root mean square error (RMSE) of the DOA, defined as

$$\text{RMSE}_\theta = \sqrt{\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\hat{\theta}_{k,i}-\theta_k\right)^2} \tag{59}$$

where n denotes the number of Monte Carlo (MC) simulation estimates, and K the number of pitches in the signal, for the resulting estimates. For comparison, we use the Cramér-Rao lower bound (CRB), the NLS estimator, and the Sub approach (see [17] for further details on these methods and for the corresponding CRB). These results have been obtained using $n = 250$ MC simulations of a single pitch signal, with $\omega_1 = 220$ Hz and $L_1 = 7$ harmonics, impinging from $\theta_1 = -30°$, where both the NLS and the Sub estimators have been allowed perfect a priori knowledge of both the number of sources and their number of harmonics, whereas the proposed method needs no such knowledge. As is clear from the figures, the HALO method offers a preferable performance as compared to the Sub estimator, and only marginally worse than the NLS estimator, in spite of both the latter being allowed perfect model orders information. Here, the number of sensors in the array was $M = 5$ and we used 20 ms of data sampled at $f_s = 8820$ Hz, i.e., $N = 176$ samples. Furthermore, $c = 343$ m/s and $d = c/f_s \approx 0.0389$ m. We proceed to consider the case of multi-pitch signals impinging on the array. Measuring as in the single-pitch case, we now form a multi-pitch signal with two pitches and fundamental frequencies $\{150,\ 220\}$ Hz containing $\{6,\ 7\}$ harmonics, coming from $\theta_1 = -30°$. Figure 5 shows the RMSE and PWL estimates, as obtained using 250 MC simulations, clearly showing that the HALO estimator is able to reach close to optimal performance also in this case. Here, no comparison is made with the NLS and Sub estimators of [17] as these are restricted to the single-pitch case. Throughout these evaluations, we have used $L_{\max} = 15$. Also, as the resulting estimates were found to be appropriately sparse when using only the convex penalties, and no reweighing steps were used. We next proceed to examine real measured signals. The measurements were made in an anechoic chamber, approximately $4 \times 4 \times 3$ meters in size, with the sensors and speakers located as shown in Figures 3 and 10. Two speakers were placed at locations (in polar coordinates) $\mathbf{s}_1 = [\theta_1, R_1] = [115.03°, 1.15\,\text{m}]$ and $\mathbf{s}_2 = [\theta_2, R_2] = [-74.53°, 1.33\,\text{m}]$, with respect to the central microphone, respectively. The positions of the sensors were determined by placing them together with the sources, using the acoustic method detailed in [41]. This is done by calibrating the sensors with a single moving source, using a correlation-based methodology. The positions were also confirmed via a computer vision approach were the positions were found by taking several photos and reconstructing the environment. The maximum deviation in position between these methods was less than 1 cm. As the spatial impulse responses of the microphones were deemed to be reasonably omni-directional, as well as roughly the same for all the microphones, no further calibration of the sensor gains were performed. The positions were then projected onto a 2-D plane using principal component analysis. In order to illustrate

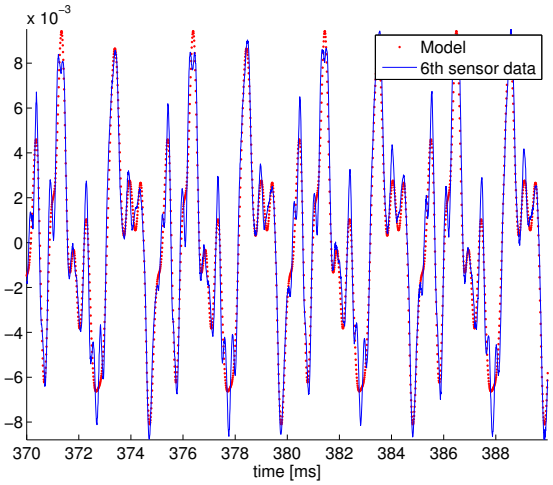

Fig. 6. Time-domain data for the 7th sensor (lined), overlaid with the signal model reconstruction (dotted).

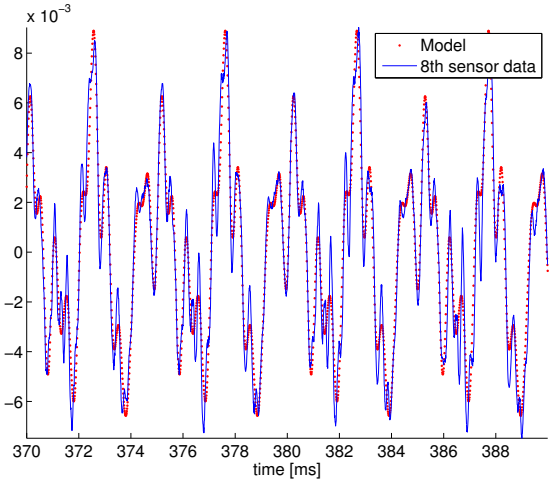

Fig. 7. Time-domain data for the 8th sensor (lined), overlaid with the signal model reconstruction (dotted).

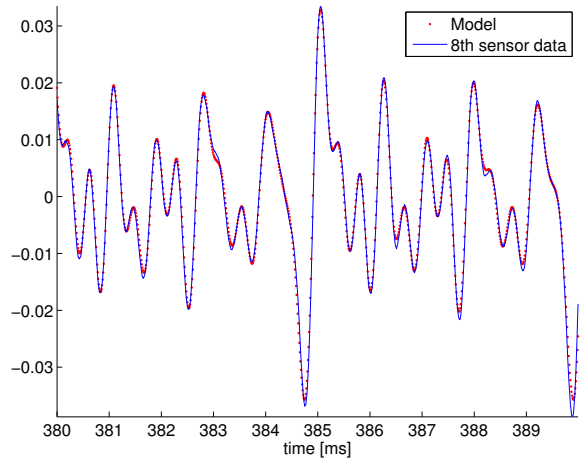

Fig. 8. Time-domain data for the 8th sensor (lined), overlaid with the signal model reconstruction (dotted).

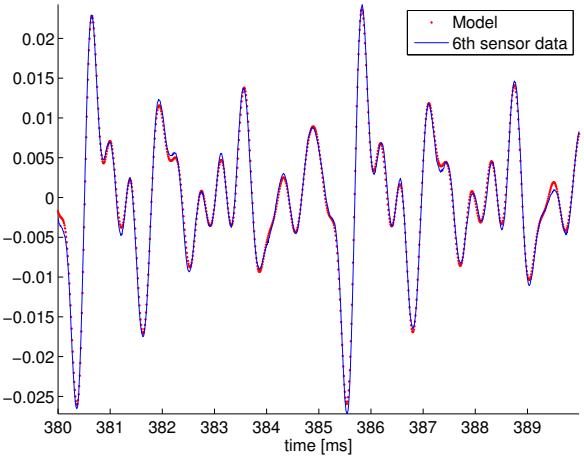

Fig. 9. Time-domain data for the 6th sensor (lined), overlaid with the signal model reconstruction (dotted).

the HALO estimator's ability to handle an environment with the same pitch signal originating from different sources, as a much simplified proof of concept for a reverberating room environment, we examine a case with two sources playing the same signal content. Both sources plays a (TIMIT) recording of a female voice saying 'Why were you away a year, Roy?', timing the source's playback so that the recording at each microphone sounds slightly echoic. The eight microphones all record at a sample rate of $f_s = 96$ kHz. The data is then divided into time frames of 10 ms, i.e., $N = 960$ samples, which allow each frame to be well modelled as being stationary. Examining a part of the speech that is voiced, arbitrarily selected as the frame starting 380 ms into the recording, about when the voice is saying the voiced phonetic sound 'a' in 'why', Figures 6 and 7 show the signal measured at the 6th and 8th microphone, respectively, together with the reconstructed signal obtained from the pitch estimation step in HALO, obtained as

$$\hat{\mathbf{Y}} = \boldsymbol{\mathcal{W}}\hat{\boldsymbol{\mathcal{B}}} \tag{60}$$

using the resulting model orders and estimates. The estimator indicate that the signal contains a single pitch at $\hat{\omega}/2\pi = 193.5$ Hz, having $\hat{L} = 12$ overtones. As is clear from the figures, the estimator is well able to model the measured signal in spite of the presence of the reverberation. Comparing the figures, one may also note the time shift between the sensors, due to the additional time-delay for the wavefront traveling between them, corresponding to a linear combination of the two sources, each with their particular TDOA and attenuation. It should also be noted that the signals are not simply time-shifted versions of each other due to the room environment and the attenuation of the signal when propagating in space (which would thus create problems for an estimator based on the cross-correlation between the sensors). The same situation is illustrated in Figures 8 and 9 showing

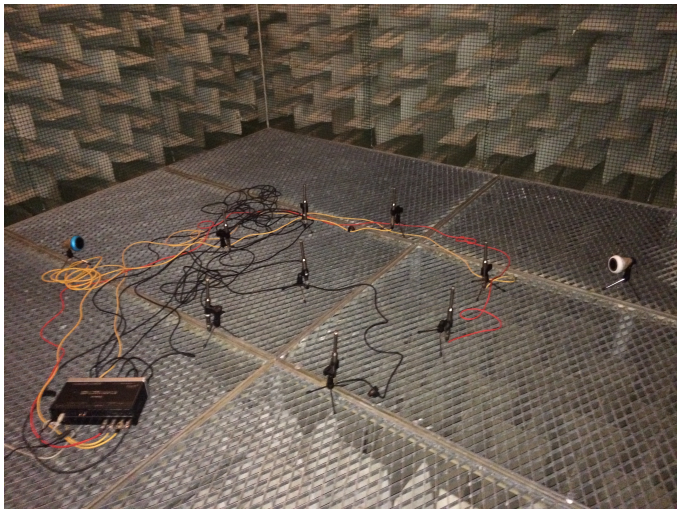

Fig. 10. A photo showing the experimental setup in the anechoic chamber.

the results when the signal source is replaced with that of a part of a (SQAM) violin signal. Again, the estimator can be seen to be able to well model the impinging signals, which is estimated as being a single pitch with the fundamental frequency $\hat{\omega}/2\pi = 198.0$ Hz, containing $\hat{L} = 14$ harmonics. In order to examine the location estimation, we construct a 2-D grid of feasible locations, chosen such that the space is discretized into 1008 points, consisting of 72 directions between $[-180°, 180°)$, spaced every $5°$, where each direction allows for ranges $R \in [0.7, 2]$ m, spaced 10 cm apart. The resulting grid is shown in Figure 11, which is roughly covering the entirety of the anechoic chamber. To also allow for far-field sources, a range of $R = \infty$ is also added to the grid for each direction, which we have chosen to illustrate by the outer circle in Figure 11. For these far-field grid points, the time-delays are instead computed as (see also [11])

$$\tau_m = \frac{\min_{\mathbf{z}} \|\mathbf{r}_m - \boldsymbol{\ell}(\mathbf{z})\|_2}{c} \tag{61}$$

for a location $\mathbf{z}$ on the line $\boldsymbol{\ell}(\cdot)$, which is perpendicular to the DOA and goes through $\mathbf{r}_1$. The figure also shows the locations for the sensors and the sound sources, as well as the estimated locations, as obtained by the second step of the HALO estimator (the estimated locations were identical for both audio recordings). The errors in position were 5 cm in range for each source, where a bias, overestimating the range, accounts for almost all of the error. On the other hand, as shown in the figure, the angles of the sources θ were accurately estimated. The overestimation of the range may to a large extent be explained by poor scaling when calibrating the array. One may note that, for localization in 3-D, the size of the dictionary will increase significantly as compared to the 2-D case used for numerical illustration in this paper. For the case above, if also the elevation angle is to be considered, having the same resolution as for the azimuth, this would yield a dictionary of 72 576 atoms. Although much larger, a sparse modeling systems of this size is by no mean impractical to work with. Also, our investigations show that a less dense location grid may be used, whereafter a zooming step can be taken. Although limited, our findings indicate that the algorithm may be used in a realistic scenarios. Clearly, further investigations are needed on more practical aspects. For future works, it is of interest to examine how the proposed method performs in more diverse scenarios. For instance, the method shows potential for use in a true reverberating environment, where instead of only one reflection, there is a multitude, giving rise to both clearly distinguishable reflections, i.e., early reverberation, and less clear such, i.e., late reverberations. Also, different combinations of microphone arrays and source signals could be evaluated, to try and find a limit for how many microphones are needed to resolve one or several sources, arbitrarily placed in the environment. Finally, the experiments shown above, as well as the mentioned suggestions, can be generalized onto a 3-D array geometry, thus adding a dimension in the localization step.

Finally, we illustrate the algorithm's performance using MC simulations, using simulated sources, one near- and one far-field source, detailed with $\omega = [200, 270]$ Hz, $L = [15, 14]$ harmonics, impinging from $\theta = [110°, -70°]$ at $R = [1.3, \infty]$ m, respectively. The sensors are placed as a uniform circular array, with 7 sensor placed evenly at a 0.5 m radius, together with a sensor being placed in the center of the array. First, we examine the position estimates using a coarse spacing for the possible sources, spaced by 11 cm in angle for all angles $\theta \in [-180°, 180°)$, and spaced by 10 cm in range, at $R \in [0.7, 3]$ m. In each MC simulation, the true location of each source was offset by a (uniformly distributed) range offset of plus minus one half the grid spacing. In all simulations, we ensured that neither of the sources were placed on a dictionary grid point. Figure 12 shows the PWL for the angle and range estimates, where the limit is chosen to be the same as the grid spacing, i.e., the ratio of estimates that are within ± 10 cm in range, and $\pm 5°$ in angle. As seen from the figure, the both the range and the DOA of the sources are well determined, indicating that even with the use of a coarse grid, one is able to obtain reliable estimates. Proceeding to instead using a fine grid, the coarse estimates may then be refined by zooming in the grid over the found locations. Using a dictionary of the same size as the coarse grid, although centered around the found estimates, yields a resolution of ± 5 mm in range and $\pm 0.25°$ in angle. Figure 13 shows the resulting RMSE for the angle and pitch estimates on the finer grid, as compared to the CRB (given in the Appendix). As can be seen from the figure, the RMSE (and the corresponding CRB) of the far-field source is somewhat lower than the near-field source, although both sources are well estimated, yielding a performance close to being optimal. The slight offset from the CRB is deemed to be largely due to a small bias in the final estimates, resulting from the smoothness of the approximative cost function resulting from the additive convex constraints. As is clear from the above presentation, the HALO estimate exploits the harmonic structure in the received audio signals to position the sources, using the pitch estimates to form

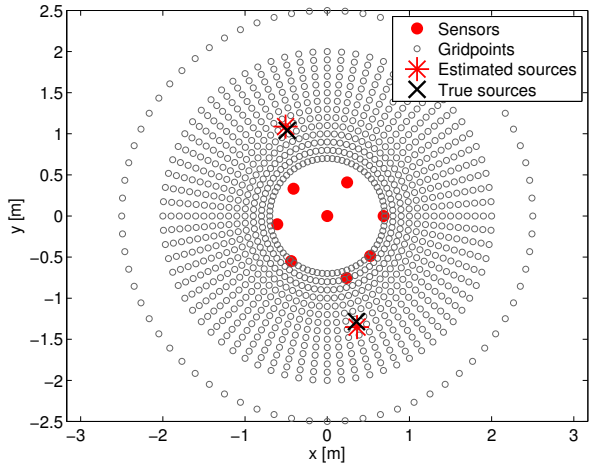

Fig. 11. The experimental setup in the anechoic chamber, showing the sensor and loudspeaker locations, the considered dictionary grid, as well as the resulting estimated as obtained by the proposed algorithm.

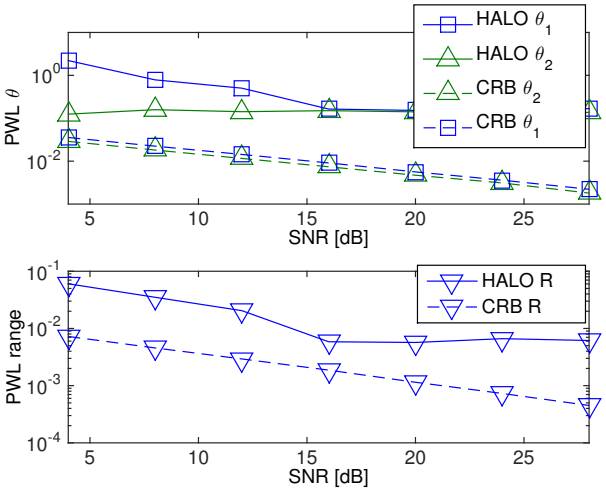

Fig. 13. The RMSE for the angle and range estimates when using a finely spaced grid, indicating the ratio of estimates that are within ± 5 mm in range, and $\pm 0.25^\circ$ in angle.

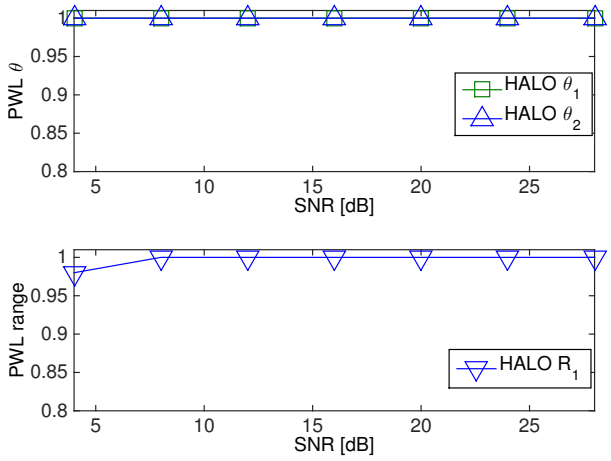

Fig. 12. The PWL ratio for the angle and range estimates when using a coarsely spaced grid, indicating the ratio of estimates that are within ± 10 cm in range, and $\pm 5^\circ$ in angle.

a sparse estimate over a wide range of feasible positions. Obviously, most audio signals are not harmonic at all times, and the estimator should thus be used in combination with a tracking technique, possibly using a methodology reminiscent to the one presented in [42], [43]. In such a tracking scheme, the estimated pitch amplitudes should be used as an indicator for the reliability of the obtained positioning, yielding poor or maybe even erroneous positioning for unvoiced or non-harmonic audio signals, whereas reasonably accurate positions may be expected for more harmonic signals.

VI. Conclusions

In this paper, we have presented an efficient sparse modeling approach for localizing harmonic audio sources using a calibrated sensor array. Assuming that each of the harmonic components in each pitch can only come from one source, the localization estimate is based on the phase and attenuation information for all of the harmonics jointly. The resulting model phases and attenuation will then depend on the source location. By using sparse modeling, the method inherently estimates both the number of sources, the number of harmonics in each source, as well as the extent of a possibly occurring reverberation. The effectiveness of the resulting algorithm is shown using both simulated and measured audio sources.

VII. Acknowledgements

The authors wish to express their gratitude to the Signal Processing Group at Electrical and Information Technology, Lund University, for allowing use of their experimental facilities, as well as to the authors of [17] for sharing their Matlab implementations.

Appendix

In this appendix, we briefly summarize the Cramér-Rao lower bound (CRB) for the examined localization problem. As is well known, under the assumption of complex circularly symmetric Gaussian distributed noise, the Slepian-Bangs formula yields [13, p. 382]

$$\left[P_{cr}^{-1}\right]_{ij} = \text{trace}\left[\mathbf{\Gamma}^{-1}\mathbf{\Gamma}_i'\mathbf{\Gamma}^{-1}\mathbf{\Gamma}_j'\right] + 2\mathcal{R}\left[\boldsymbol{\mu}_i'^H\mathbf{\Gamma}^{-1}\boldsymbol{\mu}_j'\right] \quad (62)$$

where $\mathcal{R}$ denotes the real part of a complex scalar, $\mathbf{\Gamma}$ the covariance matrix of the noise process, and $\boldsymbol{\mu}$ is the deterministic signal component, with $\mathbf{\Gamma}_i'$ and $\boldsymbol{\mu}_i'$ denoting the derivative of $\mathbf{\Gamma}$ and $\boldsymbol{\mu}$ with respect to element i of the parameter vector, respectively. For the case of uncorrelated noise with a known variance σ^2, this simplifies to

$$\left[P_{cr}^{-1}\right]_{ij} = 2\mathcal{R}\left[\boldsymbol{\mu}_i'^H\boldsymbol{\mu}_j'\right]/\sigma^2 \quad (63)$$

Using the assumed signal model as measured at sensor m, stacking the the observations as in (19), and then using the vec operator on the resulting matrix results, one obtains

the $\boldsymbol{\mu}$ function needed for the CRB calculations. Here, the parameters to be estimated are

$$\boldsymbol{\Delta} = \left\{ \left\{ a_{k,\ell}, \phi_{k,\ell} \right\}_{\ell=1,\ldots,L_k}, \omega_k, \theta_{s,k}, R_{s,k} \right\}_{\substack{s=1,\ldots,S \\ k=1,\ldots,K}} \quad (64)$$

Clearly, the resulting function may easily be derivated with respect to the magnitude, frequency and phase parameters. However, since the location parameter, $\theta_{s,k}$ and $R_{s,k}$, enter into the expression in a complicated manner depending on the sensor geometry, the corresponding derivatives are not straight forward for an arbitrary array. For this reason, for the considered array geometries, we here simply approximate the resulting expressions using numerically differentiated expressions.

REFERENCES

[1] T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "Joint DOA and Multi-Pitch Estimation Using Block Sparsity," in *39th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Florence, May 4-9 2014.

[2] J. Meldercreutz, "Om Längders Mätning Genom Dåns Tilhielp," *Vetenskapsakademiens Handlingar*, vol. 2, pp. 73–77, 1741.

[3] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148–152, Mar 1996.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberent rooms," in *Microphone Arrays: Techniques and Applications*, M. Brandstein and D. Ward, Eds., pp. 157–180. Springer-Verlag, New York, 2001.

[5] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, Nov 2003.

[6] E. Kidron, Y. Y. Schechner, and M. Elad, "Cross-Modal Localization via Sparsity," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1390–1404, April 2007.

[7] M. D. Gillette and H. F. Silverman, "A Linear Closed-Form Algorithm for Source Localization From Time-Differences of Arrival," *IEEE Signal Processing Letters*, vol. 15, pp. 1–4, 2008.

[8] K. C. Ho and M. Sun, "Passive Source Localization Using Time Differences of Arrival and Gain Ratios of Arrival," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 464–477, Feb 2008.

[9] X. Alameda-Pineda and R. Horaud, "A Geometric Approach to Sound Source Localization from Time-Delay Estimates," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1082–1095, June 2014.

[10] H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data," *Computer Speech & Language*, vol. 6, no. 2, pp. 129 – 152, 1992.

[11] H. Krim and M. Viberg, "Two Decades of Array Signal Processing Research," *IEEE Signal Process. Mag.*, pp. 67–94, July 1996.

[12] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing*, John Wiley and Sons, Inc., 2002.

[13] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, Upper Saddle River, N.J., 2005.

[14] J. Benesty, M. Sondhi, M. Mohan, and Y. Huang, *Springer handbook of speech processing*, Springer, 2008.

[15] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer-Verlag, New York, NY, 1988.

[16] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Morgan & Claypool, 2009.

[17] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 21, no. 5, pp. 923–933, 2013.

[18] S. Gerlach, S. Goetze, J. Bitzer, and S. Doclo, "Evaluation of joint position-pitch estimation algorithm for localising multiple speakers in adverse acoustical environments," in *Proc. German Annual Conference on Acoustics (DAGA)*, Düsseldorf, Germany, 2011, vol. Mar. 2011, pp. 633–634.

[19] 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, 2012.

[20] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, jul 2008, pp. 10225–10229.

[21] I. F. Gorodnitsky and B. D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, March 1997.

[22] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies, "Sparse representations of polyphonic music," *Signal Processing*, vol. 86, no. 3, pp. 417–431, March 2006.

[23] M. Genussov and I. Cohen, "Multiple fundamental frequency estimation based on sparse representations in a structured dictionary," *Digit. Signal Process.*, vol. 23, no. 1, pp. 390–400, Jan. 2013.

[24] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Estimating Multiple Pitches Using Block Sparsity," in *38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, May 26–31, 2013.

[25] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[26] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, September 1999.

[27] T. Ballal and C.J. Bleakley, "DOA Estimation of Multiple Sparse Sources Using Three Widely-Spaced Sensors," in *Proceedings of the 17th European Signal Processing Conference*, 2009.

[28] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[29] M. Elad, *Sparse and Redundant Representations*, Springer, 2010.

[30] E. J. Candès, M. B. Wakin, and S. Boyd, "Enhancing Sparsity by Reweighted l_1 Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[31] L. Qing, Z. Wen, and W. Yin, "Decentralized Jointly Sparse Optimization by Reweighted ell-q Minimization," *Signal Processing, IEEE Transactions on*, vol. 61, no. 5, pp. 1165–1170, March 2013.

[32] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Comm. Pure Appl. Math.*, vol. 63, 2010.

[33] N. R. Butt, S. I. Adalbjörnsson, S. D. Somasundaram, and A. Jakobsson, "Robust Fundamental Frequency Estimation in the Presence of Inharmonicities," in *38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, May 26–31, 2013.

[34] O. Besson and P. Stoica, "Exponential signals with time-varying amplitude: parameter estimation via polar decomposition," *Signal Processing*, vol. 66, pp. 27–43, 1998.

[35] Inc. CVX Research, "CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta," `http://cvxr.com/cvx`, Sept. 2012.

[36] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008, http://stanford.edu/∼boyd/graph_dcp.html.

[37] J. F. Sturm, "Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, August 1999.

[38] R. H. Tutuncu, K. C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Ser. B*, vol. 95, pp. 189–217, 2003.

[39] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[40] N. Parikh and S. Boyd, "Proximal Algorithms," *Found. Trends Optim.*, vol. 1, pp. 127–239, 2014.

[41] Z. Simayijiang, F. Andersson, Y. Kuang, and K. Åström, "An Automatic System for Microphone Self-Localization Using Ambient Sound," in *European Signal Processing Conference (Eusipco 2014)*, 2014.

[42] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, Sept 2004.

[43] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual Probabilistic Tracking of Multiple Speakers in Meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 601–616, Feb 2007.

Ted Kronvall (S'12) received his M.Sc. in Industrial Engineering and Management from Lund University, Sweden, in 2012. Currently, he is working towards a Ph.D. in Mathematical Statistics, as a member of the Statistical Signal Processing research group at Lund University, from where he also received the intermediate doctoral degree of Licentiate in Engineering (Lic.Eng.) in 2015. He has been a visiting researcher at the department of Systems Inovations at Osaka University, Japan. His research interest include Signal Processing for Audio, Sparse Modeling, Statistical Modeling of Spectroscopic Signals, Array Processing, and Convex Optimization.

Simon Burgess received his M.Sc. in Engineering Mathematics from Lund University, Sweden, in 2008. Between 2009 and 2011 he has been working with in industry with image analysis and teaching at Lund University. He started his Ph.D. studies in 2011 at Centre for Mathematical Sciences, Lund University, where he has been focused on calibration algorithms for audio- and radio rigs, polynomial systems, and calibration free indoor localization.

Kalle Åström (M'02) received his B.Sc in Mathematics in 1990, M.Sc. degree in Engineering Physics in 1991 and Ph.D. in Mathematics in 1996 from Lund University, Sweden. His thesis was awarded Best Nordic Ph.D. Thesis in pattern recognition and image analysis 1995-1996 at the Scandinavian Conference in Image Analysis, 1997. He has been a post-doctoral research fellow, associate professor and is now professor at the Centre for Mathematical Sciences, Lund University. His teachings include undergraduate and gradate courses in mathematics, image processing and computer vision. His current research interests include stochastic analysis of low level vision, computer vision for autonomous systems, geometry and algebra of multiple views of points, curves and surfaces, structure from sound, radio and wifi, cognitive vision, handwriting recognition, medical image analysis and bioinformatics. He is the co-author of 10 patent applications and 150 refereed conference and journal publications. He is co-founder of three spinn-off companies, Decuma, Cognimatics and Spiideo.

Stefan I. Adalbjörnsson (S'09) received his B.Sc. in Electrical and Computer Engineering from the University of Iceland in 2004, and his M.Sc. in Engineering Mathematics and Ph.D. in Mathematical Statistics from Lund University in 2009 and 2014, respectively. Currently, he is working as a post doctoral researcher in applied mathematics at the Centre for Mathematical Sciences at Lund University, working on an interdisciplinary project with the Lund University Humanities laboratory, as well as with an industry partner, Quanox, on recommendation systems and related large data problems. He has been a visiting researcher at the Spectral Analysis Laboratory in University of Florida, Gainesville. His research interest include big data analytics, recommendation systems, exploratory data analysis, and applications of sparse and convex modeling in statistical signal processing and spectral analysis.

Andreas Jakobsson (S'95-M'00-SM'06) received his M.Sc. from Lund Institute of Technology and his Ph.D. in Signal Processing from Uppsala University in 1993 and 2000, respectively. Since, he has held positions with Global IP Sound AB, the Swedish Royal Institute of Technology, King's College London, and Karlstad University, as well as held an Honorary Research Fellowship at Cardiff University. He has been a visiting researcher at King's College London, Brigham Young University, Stanford University, Katholieke Universiteit Leuven, and University of California, San Diego, as well as acted as an expert for the IAEA. He is currently Professor and Head of Mathematical Statistics at Lund University, Sweden. He has published his research findings in over 150 refereed journal and conference papers, and has filed five patents. He has also authored a book on time series analysis (Studentlitteratur, 2013), and co-authored (together with M. G. Christensen) a book on multi-pitch estimation (Morgan & Claypool, 2009). He is a member of The Royal Swedish Physiographic Society, a Senior Member of IEEE, and an Associate Editor for Elsevier Signal Processing. He has previously also been a member of the IEEE Sensor Array and Multichannel (SAM) Signal Processing Technical Committee (2008-2013), an Associate Editor for the IEEE Transactions on Signal Processing (2006-2010), the IEEE Signal Processing Letters (2007-2011), the Research Letters in Signal Processing (2007-2009), and the Journal of Electrical and Computer Engineering (2009-2014). His research interests include statistical and array signal processing, detection and estimation theory, and related application in remote sensing, telecommunication and biomedicine.