Sound Source Localization and Reconstruction Using a Wearable Microphone Array and Inertial Sensors

Clas Veibäck, Martin Skoglund, Fredrik Gustafsson and Gustaf Hendeby

Conference paper

Cite this paper as:

Veibäck, C., Skoglund, M., Gustafsson, F., Hendeby, G. Sound Source Localization and Reconstruction Using a Wearable Microphone Array and Inertial Sensors, In (eds), Proceedings of the 23rd International Conference on Information Fusion: Fusion 2020, : Institute of Electrical and Electronics Engineers (IEEE); 2020, pp. 1086-1093. ISBN: 978-0-9964527-6-2

DOI: https://doi.org/10.23919/FUSION45008.2020.9190480

Sound Source Localization and Reconstruction Using a Wearable Microphone Array and Inertial Sensors

Clas Veibäck*, Martin A. Skoglund*†, Fredrik Gustafsson*, and Gustaf Hendeby*
*Division of Automatic Control, Linköping University, SE-581 83 Linköping, Sweden
†Eriksholm Research Centre, Rørtangvej 20, DK-3070 Snekkersten, Denmark
Email: {clas.veiback,martin.skoglund,fredrik.gustafsson,gustaf.hendeby}@liu.se

***Abstract*—A wearable microphone array platform is used to localize stationary sound sources and amplify the sound in the desired directions using several beamforming methods. The platform is equipped with inertial sensors and a magnetometer allowing predictions of source locations during orientation changes and compensation for the displacement in the array configuration. The platform is modular, open and 3D printed to allow for easy reconfiguration of the array and for reuse in other applications, *e.g.*, mobile robotics. The software components are based on open source. A new method for source localization and signal reconstruction using Taylor expansion of the signals is proposed. This and various standard and non-standard Direction of Arrival (DOA) methods are evaluated in simulation and experiments with the platform to track and reconstruct multiple and single sources. Results show that sound sources can be localized and tracked robustly and accurately while rotating the platform and that the proposed method outperforms standard methods at reconstructing the signals.**

I. Introduction

Direction Of Arrival (DOA) estimation and source localization from sensor arrays have been extensively studied during the last four decades see *e.g.*, [1–4] or [5] for a recent survey. A driving application for us is Hearing Aid Systems (HAS) and several methods have been proposed in order to estimate the 3D source direction or position using HAS. In [6], a chest-worn planar microphone array is used to estimate the direction and [7] uses an array in the form of a necklace. In [8] Head-Related Transfer Functions (HRTFs) are used to estimate the source position. While tracking using DOA is important for situational awareness it is often also necessary to reconstruct source signals by beamforming for identification or presentation purposes. This is especially true in HAS in which noise should be reduced [9] and target speech needs to be amplified enabling Hearing Impaired (HI) to engage in otherwise challenging scenarios such as restaurant conversations with multiple people and strong background noise.

Classical beamforming methods often consider specific array structures, such as the Uniform Linear Array (ULA) [10] which provides a uniform spatial sampling of the wavefield. Together with the narrowband assumption this enables non-parametric narrowband DOA methods, such as MUltiple SIgnal Classification (MUSIC) [11] and Minimum Variance Distortionless Response (MVDR) [12]. Narrowband methods are size-constrained as the sensors need to be separated by at least half a wavelength of the received signal for unambiguous results. Such constraints are impractical in HAS which themself are physically constrained by their design. An option is to use differential arrays [13–15] which perform beamforming by delaying and differencing the array elements in the hardware. With differential arrays the distance between the sensors must be small enough to approximate the acoustic field pressure differentials [16].

A recent alternative for size-limited arrays is spatial delay estimation using Taylor series expansion [17, 18]. The main contribution of this paper is an extension of this method where constraints are added to enforce consistency of the estimated signals over time. To evaluate the method, a wearable microphone Array Frame (AF) with flexible configuration is developed and it contains an Inertial Measurement Unit (IMU), a magnetometer and all necessary components for computation. While not matching the form factor of hearing aids, such as binaural behind-the-ear devices, the AF is rather an idealized platform with extended capabilities. Some of these are: high-dimensional beamforming; DOA estimation in absolute coordinates; continuous beam-steering [19] with source location feedback; and easier application of experiments in ecologically valid scenarios due to its portability. Simulation and experimental results demonstrate source localization and reconstruction using the Taylor series expansion method in scenarios with single and multiple sources, and rotating AF. The Taylor series estimator is compared with several other DOA and beamformer counterparts. The source code and design files are provided as open source [1] [2].

II. Array and Source Parameters

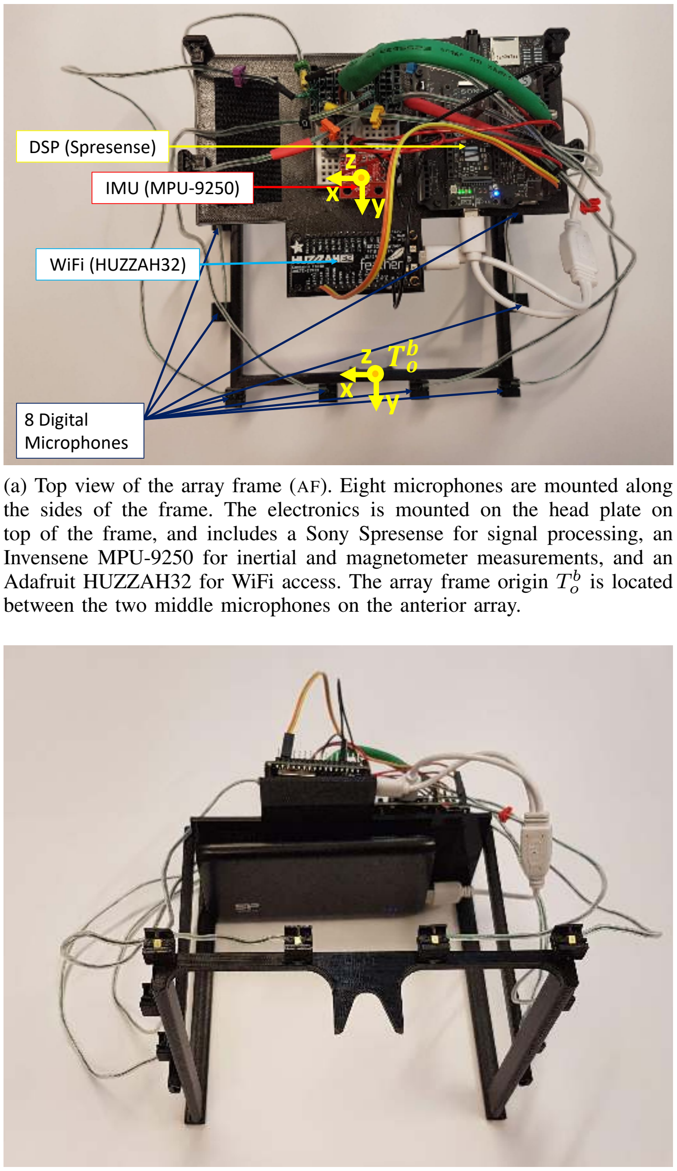
A. Array and source geometry

The AF is constructed with two two-microphone (ULAs) and one ULA with four microphones which are rigidly attached

C. Veibäck has received funding from the Oticon Foundation. G. Hendeby has received funding from the Center for Industrial Information Technology at Linköping University (CENIIT). F. Gustafsson has received funding from the Swedish Research Council through the project Scalable Kalman Filters.

[1] gitlab.liu.se/veiback-public/lindoa

[2] gitlab.liu.se/veiback-public/array-frame



(a) Top view of the array frame (AF). Eight microphones are mounted along the sides of the frame. The electronics is mounted on the head plate on top of the frame, and includes a Sony Spresense for signal processing, an Invensene MPU-9250 for inertial and magnetometer measurements, and an Adafruit HUZZAH32 for WiFi access. The array frame origin T_o^b is located between the two middle microphones on the anterior array.

(b) Front view of the array frame showing power supply mounted under the the head plate.

Fig. 1: Overview of 3D printed array frame with sensors, processing units, and power supply.

to each other, see Fig. 1, for an overview. The N sensors are distributed along these three arrays. Let the AF origin be T_o^b and the sensors expressed in the body frame, b, are denoted $T_n^b = [x_n^b, y_n^b, z_n^b]^T$, $n = 1, \dots, N$. Denote the position of sensor n in the global (inertial) frame, e, with $T_n^e = [x_n^e, y_n^e, z_n^e]^T$, then the two frames are related

$$T_n^e = \mathsf{R}^T(T_o^b + T_n^b), \tag{1}$$

where R is a rotation matrix $\{\mathsf{R} \in \mathbb{R}^{3\times 3}, \det \mathsf{R} = 1, \mathsf{R}^T = \mathsf{R}^{-1}\}$ describing the orientation of the b frame with respect to the e frame expressed in the e frame. A source in the e frame $M_i^e = [X_i^e, Y_i^e, Z_i^e]^T$ can thus be resolved at sensor n as

$$M_{in}^b = \mathsf{R}(M_i^e - T_n^e) = \mathsf{R}\, M_i^e - T_o^b - T_n^b. \tag{2}$$

B. Delay parametrization

With far-field sources the delay to sensor n with position $T_n^e = [x_n^e, y_n^e, z_n^e]^T$ is parametrized using the two angles

$$\begin{aligned} &\tau_n(\varphi_i, \theta_i) \\ &= \frac{x_n^e \sin(\varphi_i)\cos(\theta_i) + y_n^e \cos(\varphi_i)\cos(\theta_i) + z_n^e \sin(\theta_i)}{c} \\ &= \frac{1}{c}\left[\sin(\varphi_i)\cos(\theta_i), \cos(\varphi_i)\cos(\theta_i), \sin(\theta_i)\right] T_n^e \end{aligned} \tag{3}$$

where c is the speed of sound, φ_i is the azimuth angle with respect to the magnetic north and and θ_i is the elevation angle to source i with respect to the horizontal plane.

With near-field sources the delay is parametrized using the 3D (or 2D) position of the source $M_i^e = [X_i^e, Y_i^e, Z_i^e]^T$ and sensor location

$$\tau_n(M_i^e) = \frac{1}{c}\|M_i^e - T_n^e\|_2, \tag{4}$$

where $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^T\mathbf{a}}$ is the Euclidean 2-norm. It is straightforward to consider *e.g.*, unknown AF location and orientation in (3) and (4) if ego-localization is of interest or sensor position in the AF if sensor position calibration is sought.

C. Array orientation using IMU and magnetometer

The IMU, comprising a 3-axis accelerometer and a 3-axis gyroscope, is combined with the data from a 3-axis magnetometer to resolve the orientation of the AF. It is assumed that the acceleration is small compared to gravity and hence the accelerometer measurements at time k can be approximated as

$$\mathbf{y}_k^{\text{acc}} \approx \mathsf{R}_k\, \mathbf{g} + \mathbf{e}_k^{\text{acc}}, \tag{5}$$

where $\mathbf{g} = [0, 0, g]^T$ is the local gravity vector, $g \approx 9.81\text{m/s}^2$, and $\mathbf{e}_k^{\text{acc}} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{\text{acc}})$ is noise. The displacement between the IMU origin and the AF origin is also assumed negligible in the experiments. The gyroscope measurements are

$$\mathbf{y}_k^{\text{gyr}} = \boldsymbol{\omega}_k + \mathbf{e}_k^{\text{gyr}}, \tag{6}$$

where $\boldsymbol{\omega}_k$ are the angular rates of the AF and $\mathbf{e}_k^{\text{gyr}} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{\text{gyr}})$ is noise.

Similarly to the accelerometer model the magnetometer measurements are

$$\mathbf{y}_k^{\text{mag}} = \mathsf{R}_k\, \mathbf{m} + \mathbf{e}_k^{\text{mag}}, \tag{7}$$

where $\mathbf{m}$ is the local magnetic field and $e_k^{\text{mag}} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{\text{mag}})$ is noise. A convenient orientation parametrization is given by the unit quaternion [20], denoted $\mathbf{q} = [q_0\ q_1\ q_2\ q_3]^T \in \mathbb{S}^3 \subset \{\mathbb{R}^4 | \mathbf{q}^T\mathbf{q} = 1\}$ and the rotation matrix is computed from $\mathbf{q}$ as

$$\mathsf{R} = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_0q_3) & 2(q_1q_3 - q_0q_2) \\ 2(q_1q_2 - q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_0q_1) \\ 2(q_1q_3 + q_0q_2) & 2(q_2q_3 - q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix}. \tag{8}$$

The quaternion dynamics using the first order Taylor approximation and sampling interval T is

$$\mathbf{q}_{k+1} \approx (\mathbf{I} + T\mathbf{S}(\boldsymbol{\omega}_k + \mathbf{w}_k))\mathbf{q}_k = \mathbf{q}_k + T\widetilde{\mathbf{S}}(\mathbf{q}_k)(\boldsymbol{\omega}_k + \mathbf{w}_k), \tag{9}$$

where $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{\omega}) \in \mathbb{R}^3$ is process noise,

$$\mathbf{S}(\boldsymbol{\omega}) = \frac{1}{2}\begin{bmatrix} 0 & -\omega_x & -\omega_y & -\omega_z \\ \omega_x & 0 & \omega_z & -\omega_y \\ \omega_y & \omega_z & 0 & \omega_x \\ \omega_z & \omega_y & -\omega_x & 0 \end{bmatrix}, \tag{10}$$

and

$$\tilde{\mathbf{S}}(\mathbf{q}) = \frac{1}{2}\begin{bmatrix} -q_1 & -q_2 & -q_3 \\ q_0 & -q_3 & q_2 \\ q_3 & q_0 & -q_1 \\ -q_2 & q_1 & q_0 \end{bmatrix}. \tag{11}$$

See [21] for further details. To obtain estimates of the orientation the IMU currently uses a Mahony filter [22], although other filters might be more suitable in future setups.

III. Signal Models

A. Array signal model

Assuming far-field sources (planar wave and no attenuation), the N microphone signals are given by [17]

$$y^n(t) = s(t + \tau_n) + e^n(t), \quad n = 1, \dots, N, \tag{12a}$$

$$\mathbf{y}(t) = \begin{bmatrix} y^1(t) & \dots & y^N(t) \end{bmatrix}^T, \tag{12b}$$

where the delay parametrization of τ_n is omitted to keep notation easier, and $e^n(t) \sim \mathcal{N}(0, \sigma_s^2)$ is independent white noise.

B. Taylor expansion

The delayed signal in (12a) is approximated by a local Taylor series expansion [17]

$$\begin{aligned} s(t + \tau_n) &\approx \sum_{l=0}^{L} \frac{d^l s(u)}{du^l} \frac{\tau_n^l}{l!}\bigg|_{u=t} = \sum_{l=0}^{L} s^{(l)}(t) \frac{\tau_n^l}{l!} \\ &= \mathbf{h}^T(\tau_n)\mathbf{x}(t), \end{aligned} \tag{13}$$

where

$$\begin{aligned} \mathbf{x}(t) &= \begin{bmatrix} s(t) & s^{(1)}(t) & \dots & s^{(L)}(t) \end{bmatrix}^T, \\ &= \begin{bmatrix} x^0(t) & x^1(t) & \dots & x^L(t) \end{bmatrix}^T, \end{aligned} \tag{14}$$

and the vector of time delays is

$$\mathbf{h}(\tau) = \begin{bmatrix} 1 & \tau & \dots & \frac{\tau^L}{L!} \end{bmatrix}^T. \tag{15}$$

The approximation in (13) is due to the neglected higher order terms in the Taylor expansion and these errors will be included in $e^n(t)$. In this notation (12) becomes

$$y^n(t) = \mathbf{h}(\tau_n)\mathbf{x}(t) + e^n(t), \quad n = 1, \dots, N, \tag{16a}$$

$$\mathbf{y}(t) = \begin{bmatrix} y^1(t) & \dots & y^N(t) \end{bmatrix}^T = \mathbf{H}(\boldsymbol{\tau})\mathbf{x}(t) + \mathbf{e}(t), \tag{16b}$$

where $\mathbf{e}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, $\mathbf{R} = \sigma_r^2 \mathbf{I}_N$ and $\boldsymbol{\tau} = \begin{bmatrix} \tau_1 & \dots & \tau_N \end{bmatrix}^T$ is a function of the signal's direction of arrival and the geometry of the microphone array which is detailed in Section II-A.

Since the signal is sampled uniformly at times $t_k = kT$ where $k = 1, \dots, K$ are the sample indices and T is the sample time, an equivalent discrete-time notation is introduced as $\cdot_k \triangleq \cdot(t_k)$, *e.g.*, $\mathbf{y}_k \triangleq \mathbf{y}(t_k)$.

IV. Estimation

A. Least squares estimation

The array model is linear in the Taylor expansion parameters $\mathbf{x}(t)$ but nonlinear in the time delays $\boldsymbol{\tau}$. This separability can be utilized when the time delay vector $\boldsymbol{\tau}$ is given. Then $\mathbf{x}(t)$ and its covariance can be estimated using least-squares (LS) as

$$\begin{aligned} \hat{\mathbf{x}}(t) &= (\mathbf{H}^T(\boldsymbol{\tau})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\tau}))^{-1}\mathbf{H}^T(\boldsymbol{\tau})\mathbf{R}^{-1}\mathbf{y}(t) \qquad \text{(17a)} \\ &= (\mathbf{H}^T(\boldsymbol{\tau})\mathbf{H}(\boldsymbol{\tau}))^{-1}\mathbf{H}^T(\boldsymbol{\tau})\mathbf{y}(t) = \mathbf{H}^{\dagger}(\boldsymbol{\tau})\mathbf{y}(t), \\ \operatorname{cov}(\hat{\mathbf{x}}(t)) &= (\mathbf{H}^T(\boldsymbol{\tau})\mathbf{R}^{-1}\mathbf{H}(\boldsymbol{\tau}))^{-1} \\ &= (\mathbf{H}^T(\boldsymbol{\tau})\mathbf{H}(\boldsymbol{\tau}))^{-1}\sigma_r^2, \qquad \text{(17b)} \end{aligned}$$

where $\cdot^{\dagger}$ denotes the Moore-Penrose inverse. In discrete time the notation is

$$\hat{\mathbf{x}}_k = \mathbf{H}^{\dagger}(\boldsymbol{\tau})\mathbf{y}_k, \tag{18a}$$

$$\operatorname{cov}(\hat{\mathbf{x}}_k) = \mathbf{P}_k = (\mathbf{H}^T(\boldsymbol{\tau})\mathbf{H}(\boldsymbol{\tau}))^{-1}\sigma_r^2. \tag{18b}$$

This is the basis of the method denoted Linear Direction Of Arrival (LINDOA) [17].

B. Signal constraints

The Taylor series model further implies that the signal and its time derivatives in (14) are not independent between samples if the sampling interval is small. This dependence can be described by noting in (14) that

$$\dot{x}^l(t) = x^{l+1}(t), \quad l = 0, \dots, L - 1, \tag{19}$$

which in discrete time transforms to

$$x_{k+1}^l = \sum_{i=0}^{L-l} \frac{T^i}{i!} x_k^{i+l}, \quad l = 0, \dots, L - 1. \tag{20}$$

This is summarized by

$$\underline{\mathbf{I}}\mathbf{x}_{k+1} = \underline{\mathbf{F}}\mathbf{x}_k, \tag{21}$$

where $\underline{\mathbf{I}}$ is the size $L + 1$ identity matrix with the last row deleted, and $\underline{\mathbf{F}}$ is an upper-triangular Toeplitz matrix defined, using $T_l = \dfrac{T^l}{l!}$, as

$$\underline{\mathbf{F}} = \begin{bmatrix} 1 & T & T_2 & \dots & T_L \\ 0 & 1 & T & \dots & T_{L-1} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & T \end{bmatrix}. \tag{22}$$

The constraints induce coupling in the system, complicating the estimation, which can be formulated as an equality-constrained linear least-squares problem on the form

$$\hat{\mathbf{x}} = \underset{\mathbf{x}_1, \dots, \mathbf{x}_K}{\arg\min} \quad \sum_{k=1}^{K} \|\mathbf{y}_k - \mathbf{H}(\boldsymbol{\tau})\mathbf{x}_k\|^2, \tag{23a}$$

$$\text{s. t.} \qquad \underline{\mathbf{I}}\mathbf{x}_{k+1} = \underline{\mathbf{F}}\mathbf{x}_k, \quad k = 1, \dots, K - 1, \tag{23b}$$

and can be solved using methods discussed in, *e.g.*, [23]. This is the main contribution of the paper and is the basis of the method we will refer to as Time-Constrained LINDOA (TC-LINDOA). A simplification is to use inexact discretization, *e.g.*, Eulers's method, resulting in an $\underline{\mathbf{F}}$ where all T_l are replaced by zeros. Another simplification would be to add process noise to the highest derivative, $\dot{x}^L(t) = w(t)$, which in discrete time would make the equality constraints in (23b) uncertain and reduce (23) to a generalized least-squares problem [24].

C. Time delay estimation

Standard methods for estimating time delays in signals are based on finding maxima in correlation or correlation-like functions. In the Taylor series approach, using only snapshots, a search-based method is a good option. With a linear LS estimate $\hat{\mathbf{x}}_k(\boldsymbol{\tau}_k)$, explicitly depending on $\boldsymbol{\tau}_k$ (or its parametrization), the LS cost function is [17]

$$\hat{\boldsymbol{\tau}}_k = \arg\min_{\boldsymbol{\tau}_k} \|\mathbf{y}_k - \mathbf{H}(\boldsymbol{\tau}_k)\hat{\mathbf{x}}_k(\boldsymbol{\tau}_k)\|^2, \tag{24}$$

which can be solved using, *e.g.*, numerical search. For signal reconstruction the parametrization is not important but if source and array geometry is of interest the delays and its parametrization must be consistent.

With the constraints from (23) included in (24) the sequence of $\boldsymbol{\tau}$'s can be found

$$(\hat{\boldsymbol{\tau}}_1, \ldots, \hat{\boldsymbol{\tau}}_K) = \arg\min_{\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_K} \sum_{k=1}^{K} \|\mathbf{y}_k - \mathbf{H}(\boldsymbol{\tau}_k)\hat{\mathbf{x}}_k(\boldsymbol{\tau}_k)\|^2, \tag{25a}$$

$$\text{s.t. } \underline{\mathbf{I}}\hat{\mathbf{x}}_{k+1}(\boldsymbol{\tau}_{k+1}) = \underline{\mathbf{F}}\hat{\mathbf{x}}_k(\boldsymbol{\tau}_k), \quad k = 1, \ldots, K-1, \tag{25b}$$

where the constraints are now nonlinear in τ. Typically the signal variations are faster than the time delay variations and therefore several $\boldsymbol{\tau}$ may be considered equal, and thus relaxing (25).

D. State augmentation

In the case of moving sources and/or a moving AF the state can be augmented with the unknowns of the delay parametrization and the corresponding state space model is then on nonlinear form

$$\bar{\mathbf{x}}_{k+1} = \mathbf{f}(\bar{\mathbf{x}}_k) + \mathbf{w}_k, \tag{26a}$$

$$\mathbf{y}_k = \mathbf{h}(\bar{\mathbf{x}}_k) + \mathbf{e}_k. \tag{26b}$$

For instance, for far-field sources with nearly constant position and unknown AF orientation the augmented state $\bar{\mathbf{x}}_k$ consists of the signal derivatives $\mathbf{x}_k$, the orientation R_k and the directions $\{\varphi_{k,i},\ \theta_{k,i}\}$ to the sources $i = 1, \ldots, M$, and appropriate dynamic models are introduced. These models can be treated using *e.g.*, the Extended Kalman Filter (EKF) [25] or other nonlinear estimators.

E. Reconstruction

With an estimate of the parameter vector $\hat{\mathbf{x}}_k$ the signal is reconstructed as

$$\hat{s}(t_k) = \hat{x}^0(t_k) = \mathbf{1}\hat{\mathbf{x}}_k = \mathbf{h}(0)\hat{\mathbf{x}}_k, \tag{27}$$

where $\mathbf{1} = [1, 0, .., 0]$. Reconstruction of the signal at an arbitrary time $t = t_k + \tau$, is obtained as

$$\hat{s}(t) = \mathbf{h}(\tau)\hat{\mathbf{x}}_k, \tag{28}$$

where t_k is chosen to minimize $|\tau|$. The variance of the estimate is $\mathrm{var}(\hat{s}(t)) = \mathbf{h}(\tau)\mathbf{P}_k\mathbf{h}^T(\tau)$.

F. Multiple sources

By superposition, M sources are incorporated in (12a) as,

$$y^n(t) = \sum_{m=1}^{M} s_m(t + \tau_{nm}) + e^n(t), \quad n = 1, \ldots, N, \tag{29}$$

where $s_m(t)$ is the signal generated by source m and τ_{nm} is the delay between source m and sensor n. The extension to the estimation is straightforward, with an augmentation of the state vector and a set of constraints for each signal. Note, however, that the measurement matrix in (16) is not full rank for multiple sources. The signal constraints are therefore required for multiple signals to obtain a well-posed problem. A simpler alternative for a first-order model of two sources is to combine the non-differentiated states for the two sources into one state by rewriting, with sensor index n omitted,

$$\begin{aligned} y(t) &= \begin{bmatrix} 1 & \tau_1 & 1 & \tau_2 \end{bmatrix} \begin{bmatrix} s_1(t) & s_1'(t) & s_2(t) & s_2'(t) \end{bmatrix}^T \\ &= \begin{bmatrix} 1 & \tau_1 & \tau_2 \end{bmatrix} \begin{bmatrix} s_1(t) + s_2(t) & s_1'(t) & s_2'(t) \end{bmatrix}^T. \end{aligned} \tag{30}$$

The estimated derivatives can then be integrated to obtain estimates of the signals from each source. This method is denoted Differentiated LINDOA (DIFF-LINDOA).

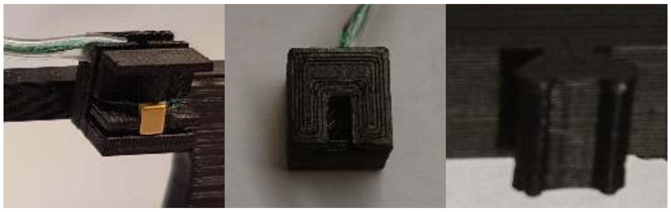
The main problem with the model in the case of multiple signals is that the number of signals and their directions of arrival need to be optimized jointly, resulting in a multi-dimensional nonlinear optimization problem. With increased dimensions, observability decreases and it becomes more difficult to find efficient numerical optimization methods. Potential modifications to lower the computational complexity would be to add layers for estimating the number of signals, their approximate directions and managing sources over time (target tracking). However, this is not an issue for reconstruction given the estimated directions to the sources.

V. System components

A. Electrical components

To reduce the development time when designing the platform, we looked for a development board with audio signal processing capabilities, interfaces for external hardware, such as sensors, memory cards and wireless connectivity, decent computational power and user-friendly development tools. Further it was desirable to keep size, weight and power consumption down to attain a portable platform with a small battery pack.

The choice landed on a Sony Spresense as the main computer, with 6 cores operating at 156 MHz, FPU, 1.5 MB SRAM and dedicated audio hardware. It has six processing cores, a variety of digital interfaces, memory card and eight digital



(a) Microphone in holder. (b) Microphone holder back with connector. (c) AF microphone connector.

Fig. 2: Microphone mount with clip-on connector system.

audio inputs which can sample up to $48\,\mathrm{kHz}$. This allows multiplexing of, *e.g.*, 16 microphones at $24\,\mathrm{kHz}$.

The microphones have amplifiers and A/D converters in the chips making the signals less prone to disturbances in the wires. To estimate the orientation of the platform, inertial and magnetometer measurements are sampled at $100\,\mathrm{Hz}$ using an Invensense MPU-9250. A power bank and WiFi access through an Adafruit HUZZAH32 make the platform completely wireless and mobile.

B. *Array frame*

The array frame was built using a 3D printer and is designed to fit a human male adult head. The microphone holders are mounted to the frame with a male-female connector, see Fig. 2, allowing for an easy change of configuration or replacements of broken components. Similar connectors are used to mount the head plate.

C. *Software*

The software is developed using the Arduino environment for Spresense with libraries available for multicore programming, audio recording and interfacing with external hardware. One core is dedicated to manage the audio recording and one core is used to interface with the IMU and run a filter for estimating orientation. The remaining cores are available for audio processing.

VI. Results

Three different scenarios are considered. The first two scenarios are a simulation and an experiment, respectively, of a stationary array frame listening to two sources, a woman and a man talking. The third scenario is an experiment of a moving AF listening to one man talking. No anechoic room was available for testing the hardware platform, so a simple scenario was setup.

A number of methods are evaluated to estimate the direction of arrival:

- Delay-And-Sum (DAS) beamforming, see *e.g.*, [26];
- First-order LINDOA, see Sec. IV-A and [17];
- First-order TC-LINDOA, see Sec. IV-B;
- Normalized Cross-Correlation NCC [10];
- Multi-Channel Cross Correlation (MCCC) [27];
- MVDR, also known as the Capon beamformer [12]; and
- MUSIC [11].

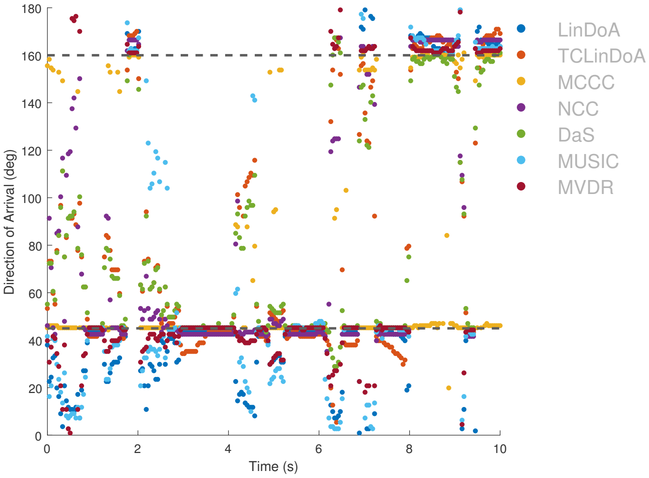


Fig. 3: Direction of arrival estimations for simulation. Each dot is an independent estimate for a time interval. 16000 samples per segment, steps of 2000 samples. The top and bottom lines are the true directions for the dialogue and monologue, respectively.

See the references for details of the methods. The last two are narrowband methods that are applied to and averaged over multiple frequencies in the range $100\,\mathrm{Hz} - 3000\,\mathrm{Hz}$.

A. *Multi-source simulation*

A man performs a monologue in Danish at 45° from the x-axis of the platform and a woman taking part in a dialogue is heard at 160°. Simulated recordings over $10\,\mathrm{s}$ are obtained for the array frame. The recordings are processed in segments of 16000 samples with an overlap of 6000 samples. All methods are employed to produce one estimate per segment, assuming a single source for each segment. However, some of the methods, *e.g.*, MCCC and TC-LINDOA are designed or can be adapted to find multiple sources simultaneously.

The results of all methods are illustrated in Fig. 3. There are natural pauses in speech, in particular in the dialogue, in which sound only arrives from one source. All methods locate the sources in such situations, while most methods produce scattered estimates when both sources are active. An exception is the MCCC approach and to a lesser extent the NCC approach, which produce very accurate direction estimates for both sources.

Although all data should be processed for accurate signal reconstruction, the direction of arrival estimations can potentially be obtained from a smaller subset of the data. The results of the proposed methods applied to a small portion of the data is shown in Fig. 4, where LINDOA is shown to produce estimates near the true directions, while TC-LINDOA is less accurate. Notice that these are independent estimates that would improve with further filtering in time.

Given an estimate of the DOA, the signals can be reconstructed. A straightforward approach is to delay and sum the recorded signals to coherently sum signals in the desired direction while cancelling noise and interfering signals. The

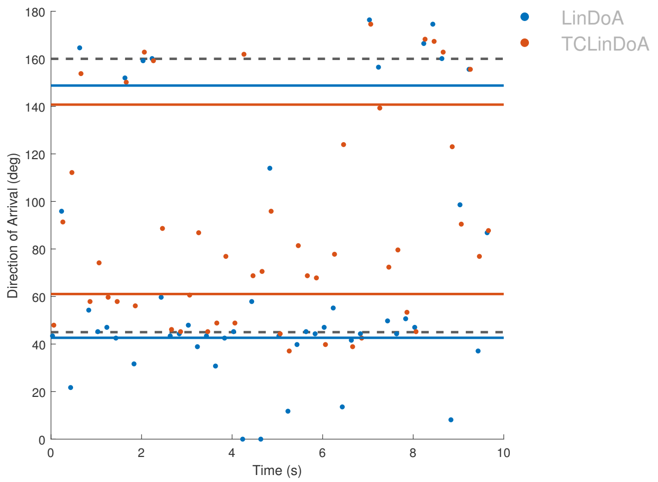


Fig. 4: Direction of arrival estimations for simulation. 48 samples per segment, steps of 9600 samples. The lines show the mean of the estimates over time.

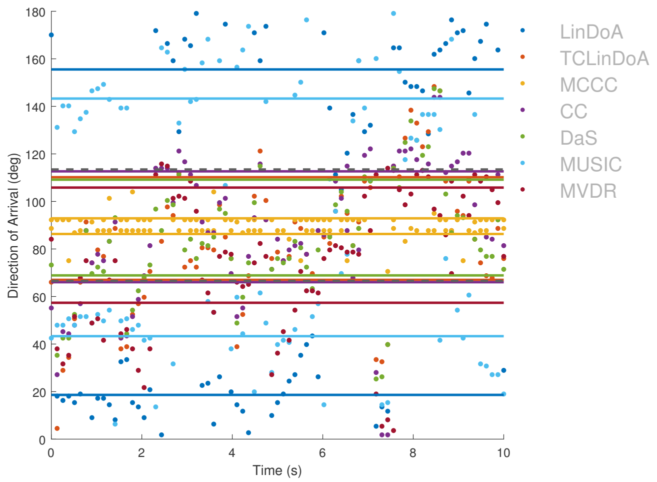


Fig. 5: Direction of arrival estimations for experiment. Each dot is an independent estimate for a time interval. 16000 samples per segment, steps of 6000 samples.

TABLE I: Correlation between the original signals and the estimated signals. Separation of the signals has been achieved by jointly estimating the signals using TC-LINDOA.

	Estimated left	**Estimated right**
True left	0.8942	0.0975
True right	0.1061	0.9534

methods LINDOA methods also generate estimates of the signals given the DOA's. Separation of sources using the true DOA is shown in Tables I, II and III, where the joint estimation using TC-LINDOA and DIFF-LINDOA perform very well. A first-order model is used for theLINDOA methods.

B. Multi-source experiment

This experiment is similar to the simulation, but is performed using speakers and the AF. The environment is rather reverberant. Seven channels are used to record the same audio used in the simulation, arriving from angles 67° and 113°. The results are shown in Fig. 5 and it is clear the performance is not on par with the simulation. Most methods are severely scattered even when one source is silent. However, some of the methods, TC-LINDOA, DAS and NCC, result in mean values that are close to the approximate true directions. The true directions are obtained geometrically by measuring the distance between the approximate location of the AF and sources.

TABLE II: Correlation between the original signals and the estimated signals. Separation of the signals has been achieved by jointly estimating the signals using DIFF-LINDOA.

	Estimated left	**Estimated right**
True left	0.8638	0.1610
True right	0.0571	0.8816

TABLE III: Correlation between the original signals and the estimated signals using Delay-and-Sum.

	Estimated Left	**Estimated Right**
True Left	0.7863	0.3114
True Right	0.5290	0.8802

Three aspects that have been ignored and are likely causes for this degradation are poor calibration, reverberation, and lack of HRTF. The microphone locations, and consequently the expected time delays, are obtained geometrically resulting in inaccurate calibration and the geometry further changes when the AF is put on the head. One solution to this problem would be to perform a calibration prior to using the AF. Another solution would be to estimate the calibration online. Estimation of a HRTF would solve this problem implicitly, while simultaneously compensating for distortions in time and frequency caused by the head. Reverberation is caused by reflections of the sound in the environment. The main difficulty is primary reflections from walls, ceiling, and floor where signals are strong and close in time with the source degrading the time delay estimation.

The reason NCC performs better than the other methods is likely that it estimates the time delays directly without regarding the parametrization of the AF. It thus maximizes the correlations of the signals and project the time delays onto the parametrization of the AF, which does not necessarily need to result in good correlation for a poorly calibrated AF.

C. Single-source experiment

This experiment tests the ability to track a single stationary source while moving the AF. The Danish monologue is recorded while rotating the AF a full turn in steps of 45°, tilting down and up and finally moving the head in arbitrary directions. Some of the methods, *e.g.*, TC-LINDOA and NCC are able to locate and track the source, which is shown in Fig. 6 and 7. The estimation results are improved by applying a moving median filter and are compared to the negative

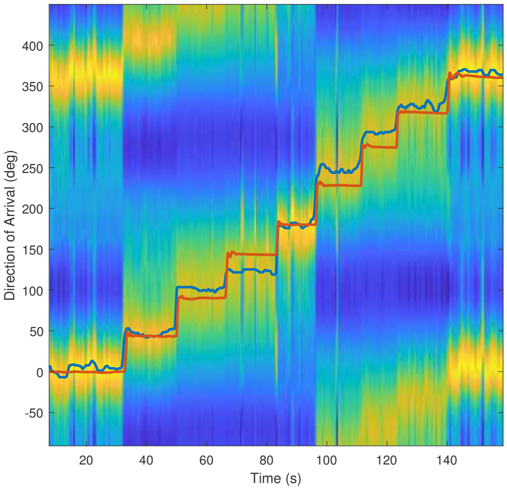


Fig. 6: Direction of arrival estimations using NCC for one source while rotating a full turn. 48000 samples per segment, steps of 16000 samples. Blue shows the estimated direction of the source relative to the AF, and red is negative yaw as measured by the IMU.

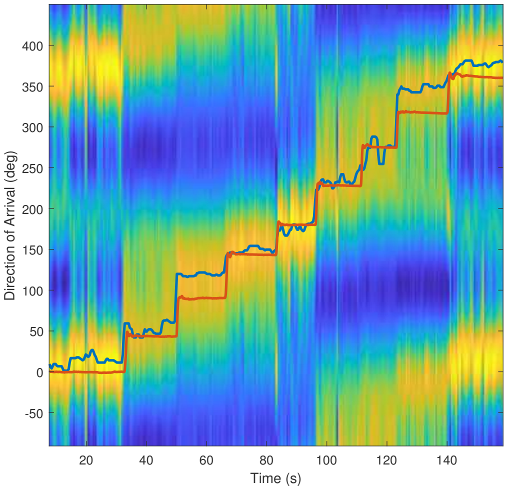


Fig. 7: Direction of arrival estimations using TC-LINDOA for one source while rotating a full turn. 48000 samples per segment, steps of 16000 samples. Blue shows the estimated direction of the source relative to the AF, and red is negative yaw as measured by the IMU.

yaw. They match reasonably, in particular when the source is directly in front or behind the AF. The method was further applied using shorter segments, which is shown in Fig. 8. The estimates are very scattered around the true direction, but would improve with a low-pass filter.

Some of the methods are unable to track the source over a full turn due to an ambiguity in the estimated direction where they cannot distinguish between sources in front and behind.

The IMU was also integrated into the algorithm for improved ability to maintain the track of a source while moving the head around. The results are shown in Fig. 9, which should be compared to the rotation in the body frame as shown in Fig. 10. The estimated direction shifts slightly at some of the movements, but generally maintains a constant estimated direction. One cause for the shifts could be translational movements, which change the direction between the AF and the source, but are not compensated for.

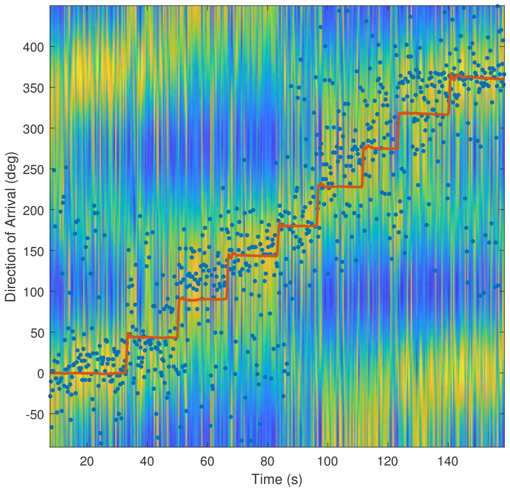


Fig. 8: Direction of arrival estimations using TC-LINDOA for one source while rotating a full turn. 480 samples per segment, steps of 9600 samples. Blue shows the estimated direction of the source relative to the AF, and red is negative yaw as measured by the IMU.

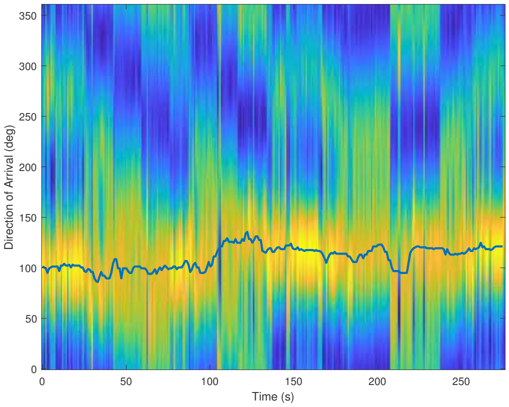


Fig. 9: The estimated direction in a global coordinate system of a stationary source as the head is moving around. The estimated direction suffers from some noise, but the IMU generally manages to compensate for the movements.

VII. Conclusions

A prototype of a mobile wearable microphone array with integrated IMU and recording capabilities was developed. The array frame has potential for further development, *e.g.*, by adding microphones, implementing online signal processing for output of filtered audio to headphones or online DOA estimation. Further, the array frame is modular such that reconfiguring the geometry or adding additional hardware

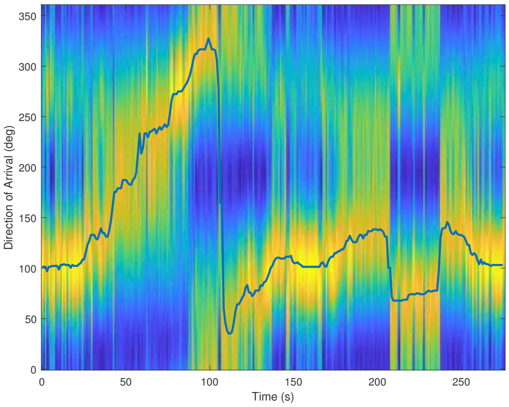


Fig. 10: The estimated direction in the body coordinate system of a stationary source as the head is moving around. This plot serves as a comparison to show the true movement.

is straightforward using a 3D printer. The prototype was simulated in an anechoic environment and used in experiments to record single and multiple sound sources in a small and rather reverberant room.

The LINDOA algorithm was extended to account for temporal constraints of the signals which we call TC-LINDOA, allowing estimation of the location of multiple sources as well as reconstruction of the original signals. The algorithm was compared to several other methods, and while location estimation performance was barely on par with existing methods, the reconstruction of the original signals was shown to be superior to the other methods. The advantage of integrating an IMU into the AF was also demonstrated in an experiment with a single source.

The first step in future work is to apply filtering versions of TC-LINDOA by introducing process noise. In a filtering approach moving sources could be handled in a target tracking fashion. Effort should also be on putting more computations onto the platform allowing for real-time processing and *e.g.*, output reconstructed audio to headphone stereo pair passed through HRTF or simply angular based stereo delay indicating sound source direction. To improve performance the AF should be calibrated by means of HRTF or similar, and the empirical and theoretical directional gains should be studied. The use of inertial sensors and a magnetometer together with 3D source tracking further opens up for increased situational awareness *e.g.*, mapping and characterizing sources and reverberations.

References

[1] H. Krim and M. Viberg, “Two decades of array signal processing research: the parametric approach,” *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[2] P. Stoica, P. Babu, and J. Li, “SPICE: A sparse covariance-based estimation method for array processing,” *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 629–638, Feb. 2011.

[3] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, “Informed TDoA-based direction of arrival estimation for hearing aid applications,” in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec. 2015, pp. 953–957.

[4] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[5] C. Rascon and I. Meza, “Localization of sound sources in robotics: A review,” *Robotics and Autonomous Systems*, vol. 96, pp. 184 – 210, 2017.

[6] W.-C. Wu, C.-H. Hsieh, H.-C. Huang, O. T. C. Chen, and Y.-J. Fang, “Hearing aid system with 3D sound localization,” in *TENCON 2007 - 2007 IEEE Region 10 Conference*, Oct. 2007, pp. 1–4.

[7] B. Widrow, “A microphone array for hearing aids,” *IEEE Circuits and Systems Magazine*, vol. 1, no. 2, pp. 26–32, Feb. 2001.

[8] F. Keyrouz, “Advanced binaural sound localization in 3-d for humanoid robots,” *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sep. 2014.

[9] J. Jensen and M. S. Pedersen, “Analysis of beamformer directed single-channel noise reduction system for hearing aid applications,” in *Proc. Int. Conf. Acoust., Speech, Signal Processing,*. Brisbane, Australia: IEEE, Apr. 2015, pp. 5728–5732.

[10] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005.

[11] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[12] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[13] G. W. Elko and Anh-Tho Nguyen Pong, “A steerable and variable first-order differential microphone array,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Munich, Germany, Apr. 1997, pp. 223–226 vol.1.

[14] G. W. Elko and J. Meyer, “Second-order differential adaptive microphone array,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 73–76.

[15] M. Buck and M. Rößler, “First order differential arrays for automotive applications,” in *7th International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, 2001.

[16] E. De Sena, H. Hacihabiboglu, and Z. Cvetkovic, “On the design and implementation of higher order differential microphones,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 162–174, Jan. 2012.

[17] F. Gustafsson, G. Hendeby, D. Lindgren, G. Mathai, and H. Habberstad, “Direction of arrival estimation in sensor arrays using local series expansion of the received signal,” in *2015 18th International Conference on Information Fusion (Fusion)*, Washington, DC, USA, Jul. 2015, pp. 761–766.

[18] G. Hendeby, T. Lunner, and F. Gustafsson, “Direction of arrival estimation using local series expansion evaluated on acoustic data,” *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2522–2522, Oct. 2017.

[19] A. Bernardini, M. D’Aria, R. Sannino, and A. Sarti, “Efficient continuous beam steering for planar arrays of differential microphones,” *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 794–798, Jun. 2017.

[20] S. W. Hamilton, “On quaternions; or on a new system of imaginaries in algebra,” *Philosophical Magazine*, vol. xxv, pp. 10–13, Jul. 1844.

[21] M. A. Skoglund, Z. Sjanic, and M. Kok, “On Orientation Estimation using Iterative Methods in Euclidean Space,” in *Proceedings of the 20th International Conference on Information Fusion (FUSION)*, Xi’an, China, Jul. 2017.

[22] R. Mahony, T. Hamel, and J. Pflimlin, “Nonlinear complementary filters on the special orthogonal group,” *IEEE Transactions on Automatic Control*, vol. 53, no. 5, pp. 1203–1218, Jun. 2008.

[23] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge University Press, 2018.

[24] F. Gustafsson, *Statistical Sensor Fusion*. Utbildningshuset/Studentlitteratur, 2012.

[25] G. L. Smith, S. F. Schmidt, and L. A. McGee, “Application of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle,” National Aeronautics and Space Administration, Washington D. C., USA, Tech. Rep. NASA TR R-135, 1962.

[26] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Wiley, 2002.

[27] D. Cochran, H. Gish, and D. Sinno, “A geometric approach to multiple-channel signal detection,” *IEEE Transactions on Signal Processing*, vol. 43, no. 9, pp. 2049–2057, Sep. 1995.