Data-Driven Multi-Microphone Speaker Localization on Manifolds

Other titles in Foundations and Trends® in Signal Processing

Recent Advances in Clock Synchronization for Packet-Switched Networks
Anantha K. Karthik and Rick S. Blum
ISBN: 978-1-68083-726-1

Biomedical Image Reconstruction: From the Foundations to Deep Neural Networks
Michael T. McCann and Michael Unser
ISBN: 978-1-68083-650-9

Compressed Sensing with Applications in Wireless Networks
Markus Leinonen, Marian Codreanu and Georgios B. Giannakis
ISBN: 978-1-68083-646-2

Data-Driven Multi-Microphone Speaker Localization on Manifolds

Bracha Laufer-Goldshtein
Faculty of Engineering
Bar-Ilan University
Israel
Bracha.Laufer@biu.ac.il

Ronen Talmon
Viterbi Faculty of Electrical Engineering
The Technion-Israel Institute of Technology
Israel
ronen@ee.technion.ac.il

Sharon Gannot
Faculty of Engineering
Bar-Ilan University
Israel
Sharon.Gannot@biu.ac.il

the essence of knowledge
Boston — Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

B. Laufer-Goldshtein, R. Talmon and S. Gannot. *Data-Driven Multi-Microphone Speaker Localization on Manifolds*. Foundations and Trends® in Signal Processing, vol. 14, no. 1–2, pp. 1–161, 2020.

ISBN: 978-1-68083-737-7

Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
  - Classification and detection
  - Estimation and regression
  - Tree-structured methods

Information for Librarians

Foundations and Trends® in Signal Processing, 2020, Volume 14, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Contents

1 Background **3**
1.1 Room Acoustics . . . 4
1.2 Classical Localization and Tracking . . . 7
1.3 Data-Driven Localization and Tracking . . . 10
1.4 Manifold-Based Localization and Tracking . . . 12
1.5 Outline of Monograph . . . 14

2 Mathematical Foundations **18**
2.1 Manifold Learning . . . 19
2.2 Locally-Linear Embedding . . . 21
2.3 Isomap . . . 22
2.4 Diffusion Maps Formulation . . . 23
2.5 Manifold Regularization in Reproducing Kernel Hilbert Spaces . . . 36

3 Data Model and Acoustic Features **44**
3.1 Data Model . . . 44
3.2 Feature Extraction . . . 45
3.3 Training and Test Sets . . . 48

4 From High-Dimensional Representation to Low-Dimensional Manifold **49**
4.1 The Acoustic Manifold 50
4.2 Distance Measures 51
4.3 Analysis of the Manifold 53

5 Data-Driven Source Localization: A Single Microphone Pair **59**
5.1 Localization Based on Diffusion Maps 60
5.2 Localization Based on Manifold Regularization 64
5.3 Results – Simulated Data 68
5.4 Results – Real Recordings 77

6 Bayesian Perspective **84**
6.1 Bayesian Inference with Gaussian Process 84
6.2 Data-Driven Prior 87
6.3 Analogy to Manifold Regularization for Localization 88

7 Data-Driven Source Localization: Ad Hoc Array **91**
7.1 Manifold-Based Gaussian Process 91
7.2 Multi-Node Data Fusion 95
7.3 Alternating Diffusion Interpretation 98
7.4 Bayesian Inference with Multiple-Manifold Gaussian Process 99
7.5 Adaptive Algorithm 99
7.6 Learning the Hyperparameters 101
7.7 Computational Complexity 102
7.8 Experimental Results 105

8 Data-Driven Speaker Tracking **118**
8.1 Problem Formulation 119
8.2 Manifold-Based Propagation Model 120
8.3 TDOA Observation Model 124

8.4 Extended Kalman Filter Tracking 125
8.5 Experimental Study 128

9 Summary and Future Directions **138**
9.1 Summary 138
9.2 Future Research Directions 140

References **142**

Data-Driven Multi-Microphone Speaker Localization on Manifolds

Bracha Laufer-Goldshtein[1], Ronen Talmon[2] and Sharon Gannot[3]

[1]*Faculty of Engineering, Bar-Ilan University, Israel; Bracha.Laufer@biu.ac.il*
[2]*Viterbi Faculty of Electrical Engineering, The Technion-Israel Institute of Technology, Israel; ronen@ee.technion.ac.il*
[3]*Faculty of Engineering, Bar-Ilan University, Israel; Sharon.Gannot@biu.ac.il*

ABSTRACT

Speech enhancement is a core problem in audio signal processing with commercial applications in devices as diverse as mobile phones, conference call systems, smart assistants, and hearing aids. An essential component in the design of speech enhancement algorithms is acoustic source localization. Speaker localization is also directly applicable to many other audio related tasks, e.g., automated camera steering, teleconferencing systems, and robot audition.

From a signal processing perspective, speaker localization is the task of mapping multichannel speech signals to 3-D source coordinates. To obtain viable solutions for this mapping, an accurate description of the source wave propagation captured by the respective acoustic channel is required. In fact, the acoustic channels can be considered as the spatial *fingerprints* characterizing the positions of each of the sources in a reverberant enclosure. These fingerprints represent complex reflection patterns stemming from the surfaces and objects characterizing the enclosure. Hence, they are

Bracha Laufer-Goldshtein, Ronen Talmon and Sharon Gannot (2020), “Data-Driven Multi-Microphone Speaker Localization on Manifolds”, Foundations and Trends® in Signal Processing: Vol. 14, No. 1–2, pp 1–161. DOI: 10.1561/2000000098.

usually modelled by a very large number of coefficients, resulting in an intricate high-dimensional representation.

We claim that in static acoustic environments, despite the high dimensional representation, the difference between acoustic channels can be attributed mainly to changes in the source position. Thus, the true intrinsic dimensionality of the variations of the acoustic channels are significantly smaller than the number of variables commonly used to represent them; that is, the acoustic channels pertain to a low-dimensional manifold that can be inferred from data using nonlinear dimensionality reduction techniques. A comprehensive experimental study carried out in a real-life acoustic environment demonstrates the validity of the proposed manifold-based paradigm.

Motivated by this result, several high-performance localization and tracking methods were developed by harnessing novel mathematical tools for learning over manifolds, including diffusion maps, semi-supervised learning, optimization in reproducing kernel Hilbert spaces and Gaussian process inference. We present two localization algorithms that were designed for a single microphone array of two microphones. These algorithms were extended to several distributed arrays by merging the information of the different manifolds associated with each array. Tracking a moving source was also addressed by a data-driven propagation model relating movements on the abstract manifold to the actual source displacements. This data-driven propagation model was combined with a classical localization approach, in a hybrid algorithm that ties together the two worlds of classical and data-driven localization, while gaining the benefits of both. We show that the proposed algorithms outperform state-of-the-art localization methods, and obtain high accuracy in challenging noisy and reverberant environments.

1

Background

Acoustic source localization is an essential component in various audio applications, such as automated camera steering and teleconferencing systems [65], speaker separation [101], robot audition [53, 62, 114, 121, 157] and drone audition [160]. For example, smart speakers require localization capabilities in order to determine the speakers in the scene and their role. Based on the location information, they can construct a direct-path steering vector to enhance the desired speaker. They may also carry out location specific tasks, such as switching the lights on and off, steering a camera, etc.

Driven by its wide applicability, the localization problem has attracted significant research attention, resulting in the development of a large variety of localization methods during the last few decades. Nevertheless, the main challenge still facing the research community is to achieve robust localization in adverse conditions, namely, in the presence of background noise and reverberations, which are the main factors in the performance degradation of localization algorithms.

In recent years, the main paradigm in localization research was based on physical models that rely on certain assumptions regarding the propagation model and the statistics of the source signal and the noise.

However, for real-world scenarios, characterized by complex reflection patterns, intricate descriptive models are required, which are difficult to estimate. Recently, the interest in applying learning-based localization approaches has been growing. Typically, these approaches assume that a training set of prerecorded measurements is given in advance. Based on this training data, they attempt to learn the characteristics of the acoustic environment directly from the data rather than using a predefined physical model.

1.1 Room Acoustics

Acoustic source localization is the task of recovering the coordinates of an acoustic source based on the signals measured in an array of microphones. Estimating only the direction of the source with respect to the array is referred to as the direction of arrival (DOA) estimation. In free-field (anechoic) environment and assuming far-field conditions, the signal measured by a microphone is a delayed version of the sound wave emitted by the source. For a uniform linear array (ULA), the time difference of arrival (TDOA) with respect to a reference microphone is geometrically related to the source DOA.

In an enclosure, the source sound propagates along multiple acoustic paths including the direct-path propagation as well as the reflections from the surfaces defining the enclosure, e.g., walls, floor, ceiling and objects, what is known as *reverberation*. As a result, the signal measured in the microphone can be expressed as the convolution between the source signal and the acoustic impulse response (AIR) relating the source and the microphone. Typical AIR consists of hundreds of taps that can be divided into three major parts: the direct path, the early reflections and the late reflections. While the early reflections correspond to the first few reflections and are sparsely distributed over time, the late reflections are highly dense and form an exponentially decreasing tail. An illustration of a typical AIR is given in Figure 1.1. An example of an AIR recorded at the Bar-Ilan University (BIU) acoustic lab with reverberation time of 610 ms and drawn from the database in [58] is given in Figure 1.2.

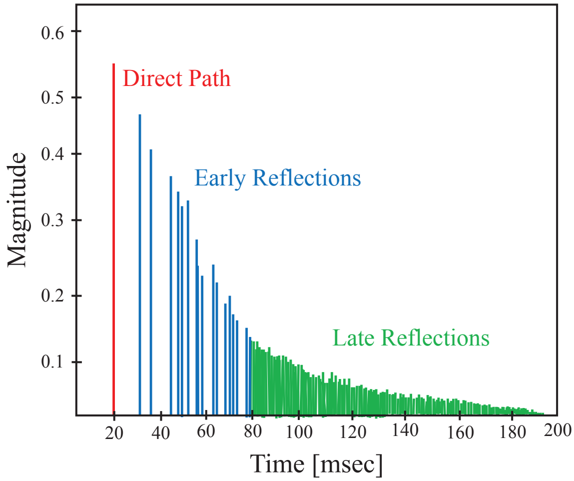

Figure 1.1: Illustration of a typical room impulse response in a reverberant environment.

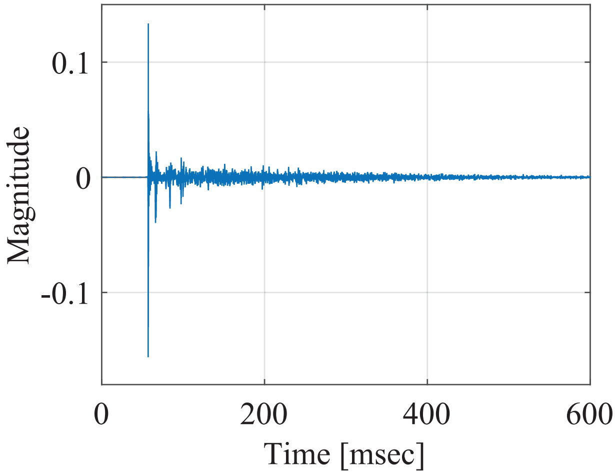

Figure 1.2: Example of room impulse response recorded at the BIU acoustic lab with reverberation time of 610 ms.

The reflections can be modeled using the image source model (ISM) by an infinite series of image sources located in mirrored rooms expanding in all three dimensions [4, 119]. However, the late reflections part does not have a distinct directionality since it consists of a superposition

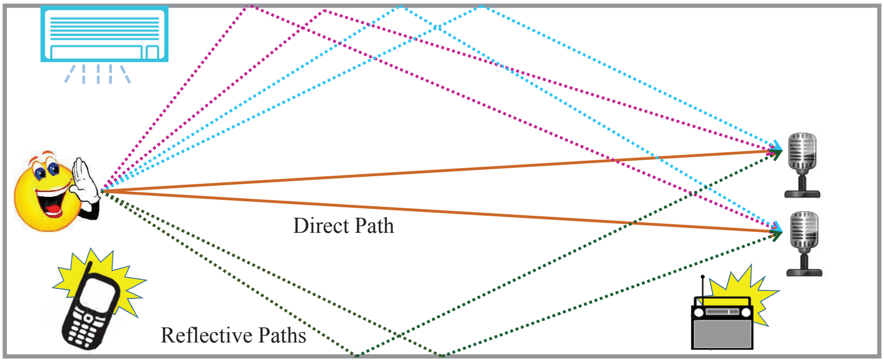

Figure 1.3: Illustration of a typical acoustic environment.

of thousands of reflections, and therefore can be statistically modeled using the law of large numbers as a zero-mean Gaussian noise signal with a decaying amplitude [120]. When the reverberation time is high, the late reflections can be modeled as a diffuse, homogeneous and isotropic field, which power is equal in all directions [29, 54].

The presence of reverberation complicates the localization task since the sound comes from many directions at the same time. In many real-life scenarios, noise sources, such as those stemming from electronic devices, air-conditioning systems and traffic, are usually present and affect the quality of the measured microphone signals and the ability of localizing the desired source. An illustration of a typical noisy and reverberant acoustic environment is given in Figure 1.3.

One important application of speaker localization is in the domain of beamformer design. Beamformers are spatial filters applied to multichannel measurements and are widely used in *speech enhancement* applications, namely to obtain noise reduction, dereverberation or separation of several mixed sources. Beamforming is obtained by multiplying the measured microphone signals by a weight function and then summing them together (usually per frequency bin). The weights of the beamformer are designed to utilize the spatial diversity of the different sound components and the noise, namely, that they come from different directions. A common spatial filter is the delay-and-sum (DS) beamformer, whose weights compensate the delay differences between the microphones, and hence require the knowledge or the estimation of the TDOAs associated with the desired source. This way, the output of the beamformer is focused on the desired source while minimizing noise and reverberation arriving from other directions. DOA estimates are also utilized for more sophisticated beamforming and separation schemes [104, 155].

Conventional beamformers that are built on the basis of the direct sound only treat the reflections as interference, and hence neglect a major part of the sound energy. They also ignore the correlation between the direct sound and its reflections, which may result in a distorted output. In [48, 102] it was shown that utilizing the entire acoustic propagation path, may significantly improve the performance of speech enhancement algorithms. Since the dimension of the full propagation path is very high, it may result in higher spatial resolution and better separation capabilities, even with a small number of microphones. For example, it can be used to extract sources with the same line-of-sight, which is impossible for beamformers that are based solely on the DOA [58]. This observation motivates the use of the full acoustic propagation path also for the localization task, as adopted by the methods presented in this monograph.

1.2 Classical Localization and Tracking

Classical localization methods usually focus on the direct path only, and ignore or mitigate the reflective part. Traditional localization methods can be broadly divided into three main categories: methods based on the maximization of the steered response power (SRP) of a beamformer output, high-resolution spectral estimation techniques, and dual-stage approaches that rely on a TDOA estimation. In the first category, the position is estimated directly from the measured signals after they have been filtered and summed together. Commonly, the maximum likelihood (ML) criterion is applied, which in the case of a single source leads to searching the maxima of the output power of a beamformer steered to different locations [171]. The second category consists of high resolution methods, such as the multiple signal classification (MUSIC) [137] and estimation of signal parameters via rotational invariance (ESPRIT) [126] algorithms, which are based on the spectral analysis of the correlation matrix of the measured signals. Subspace methods can also be applied using spherical harmonics [1, 110, 154]. In the third category, a dual stage approach is applied. In the first stage, the TDOAs of different pairs of microphones are estimated and collected. The different TDOA readings correspond to single-sided hyperbolic hyperplanes (in 3D) representing possible positions. In the second stage, the geometric intersection of these hyperplanes is recovered, which yields the estimated position [17, 65, 135]. In these dual-step approaches, the quality of the localization is strongly dependent on the quality of the TDOA estimation in the first stage.

The classical method for TDOA estimation is the generalized cross-correlation (GCC) algorithm introduced by Knapp and Carter in their landmark paper [73]. Many improvements on the GCC method for reverberant environments were proposed, e.g., in [16, 42, 127, 136, 145]. Among these methods for TDOA estimation in reverberant conditions, there are subspace methods based on adaptive eigenvalue decomposition [11] and generalized eigenvalue decomposition [38]. Of special importance is the steered response power phase transformation (SRP-PHAT) algorithm proposed in [34]. This method is related to both the first and third categories, since it combines in a single step the features of a steered beamformer with those of the phase transform weighting of the GCC algorithm.

Localization capabilities can be enhanced using model-based methods, assuming certain structures of either the speech signal or the acoustic channels. In the study of [36], an autoregressive (AR) modeling for the speech components was used, and in [66, 67] the sources were modeled as sums of harmonically related sinusoids, which describe many musical instruments and voiced speech. A model for the early reflections of the acoustic channels, based on which the early reflections were iteratively estimated, was presented in [68].

In tracking scenarios, the source is moving in the enclosure in a continuous trajectory, implying that source positions in successive time steps are related. Bayesian inference approaches, which model the varying source position as a stochastic process, are widely used. These methods commonly rely on estimated TDOAs, leading to nonlinear and non-Gaussian models, which can be solved, using, for example, the unscented Kalman filter, the extended Kalman filter (EKF) [47], and particle filters [97, 158, 162].

In real environments, the presence of noise or reverberations frequently yields unreliable observations with spurious peaks, which may lead to severe performance degradation. Several attempts to mitigate the harmful effect of noise and reverberations were made. In [175], an extended particle filter (EPF) solution was proposed, where an EKF is used to derive an optimal importance function for a particle filter. A multiple-hypothesis model accounting for the multipath nature of the sound propagation in reverberant enclosures was presented in [158], and a combination of this model with an EPF was presented in [92]. In [6, 43], a tracker was proposed based on a probability hypothesis density (PHD) filter, which is a first moment approximation of the target probability density. Robust tracking methods that use special array constellations were also proposed, such as spherical microphone arrays [77] and distributed networks [174, 176]. In [156], a robust tracker based on a distributed unscented Kalman filter was proposed, in which an interacting multiple model [15] is used for accommodating the different possible motion dynamics of the speaker, yielding a smoothed trajectory of the speaker's movement in noisy and reverberant environments. Distributed acoustic tracking that incorporates the coherent-to-diffuse ratio as a measure of DOA reliability was proposed in [44]. An additional approach for enhancing the localization robustness is to fuse several observation modalities, as demonstrated in audio-visual tracking methods [51, 149, 177, 178].

Localization and tracking of multiple speakers have also been widely investigated. Many approaches rely on the W-disjoint property of the speech signal in the short-time Fourier transform (STFT) domain [172], namely, that each time-frequency (TF) bin is dominated by a single speaker. In [99], an SRP estimate of the source position is obtained for each TF bin, and the different estimates are clustered to the different speakers using a mixture of Gaussians (MoG) model. In [100], an MoG model was proposed, in which the centroids of the different Gaussians in the mixture are associated with a grid of candidate source positions. Using Expectation-Maximization (EM) iterations, the TF bins are clustered to the different Gaussians, and the locations of the sources are estimated by selecting the Gaussians with the largest number of TF bins associations. The algorithm was extended to multiple-speaker tracking using two recursive EM variants in [139]. A further extension to distributed networks was proposed in [40]. Several improvements in noisy and reverberant conditions were presented in [41, 94, 163, 164].

We conclude that in the adverse conditions of noise and reverberation the capabilities of most of the classical localization and tracking approaches are limited. The main problem is that the reverberant nature of real-world acoustic scenarios leads to intricate acoustic channels with complex reflection patterns. Only approximated models, relying on some predefined statistical or physical assumptions, exist, which are unable to describe the acoustic channels comprehensively. In the presence of noise and reverberation, inaccurate modeling and model estimation errors frequently result in a degraded localization and tracking performance.

1.3 Data-Driven Localization and Tracking

Learning-based approaches have been proposed for both microphone array and binaural localization. In the binaural hearing context, Deleforge and Horaud proposed a probabilistic piecewise affine regression model that infers the localization-to-interaural data mapping and its inverse [31]. They extended this approach to the case of multiple sources using the variational EM framework [32, 33]. In [106], another approach was presented based on a Gaussian Mixture Model (GMM), which was used to learn the azimuth-dependent distribution of the binaural feature space. In [167], a binaural localization method was proposed in which the mutual information between each of the spatial cues and the corresponding source location is assessed. A method for DOA estimation of multiple sources using an EM clustering approach was presented in [169]. A method for localizing a source positioned behind an obstacle blocking the direct propagation path was presented in [72]. The algorithm uses co-sparse data analysis based on the physical model of the wave propagation. The model was extended in [14] to the case where the physical properties of the enclosure are not known in advance.

Recently, an increasing effort has been made to adopt deep neural networks (DNN) models for supervised localization using various network architectures and different type of input features [30, 95, 96, 170]. In the study in [168], GCC-based feature vectors were extracted and used for training a multilayer perceptron neural network, whose output is the source DOA. The eigenvectors of the spatial correlation matrix served as input features in [147] for a hierarchical network that integrates sub-band information for single-speaker localization. An extension to multi-speaker localization was presented in [146], and an adaptation mechanism that resolves the problem of mismatch in training and test characteristics was derived in [148]. In [23], a convolution neural network (CNN) based classification method for broadband DOA estimation was proposed, where the phase component of the short-time Fourier transform coefficients of the received microphone signals was directly fed into the CNN. The assumption of disjoint speaker activities was utilized in [24] to train a CNN using synthesized noise signals for multi-speaker localization. A likelihood-based encoding of the network output, which naturally allows the detection of an arbitrary number of sources, was presented in [60]. In [2], a convolution and recurrent neural network (CRNN) was proposed for estimating the DOA of multiple sources. No explicit feature extraction step is performed, as the magnitudes and phases of the spectrograms of all the channels are used directly as input to the network. For the same task, a simpler CRNN architecture that utilizes acoustic intensity features as inputs was proposed in [118].

Speaker localization can be utilized in multichannel ASR systems that commonly has a two-stage processing step of speech enhancement, including localization, beamforming and postfiltering, and acoustic modeling. In a recent line of works [93, 128–133] it was proposed to apply the multichannel enhancement jointly with acoustic modeling in a deep neural network framework. In [128] it was proposed to use the raw waveforms, rather than the log-mel features, for a single-channel speech recognition task. In the proposed architecture, the first layer is a time-convolutional layer, which can be thought of as a filterbank followed by a nonlinearity, and the output of this layer is passed to a Convolutional, Long Short-Term Memory Deep Neural Network (CLDNN) that learns the acoustic model. This method was extended to a multichannel setting in [129], showing that the proposed architecture learns to apply spatial filtering and outperforms delay-and-sum beamformer constructed with the true TDOAs. Additional improvements for this model were presented in [93, 130–133].

Most of the DNN-based localization approaches are formulated as a classification problem designed to produce a quantized estimate at a predefined grid of fixed locations. However, when addressed as a continuous regression problem, localization accuracy can be improved. An additional major problem of DNN-based localization methods is that they require a large amount of training data, the acquisition of which is frequently very difficult and time-consuming. In addition, these methods are highly prone to overfitting, and there is no guarantee they can generalize well to different acoustic scenarios beyond that used during training.

1.4 Manifold-Based Localization and Tracking

In this monograph we present a novel family of localization and tracking methods. As opposed to classical localization methods that usually ignore the richness of the acoustic propagation path, the methods presented here represent a new paradigm, in which the full intricate reflection patterns are utilized. This way we show that the intricate acoustic reflection patterns define a *fingerprint*, uniquely characterizing the source location in the enclosure. To deal with the complexity of the acoustic propagation we harness the power of manifold learning, which explores structures in high-dimensional data and extract simplified informative representations that capture its controlling parameters. We will show that the collection of acoustic fingerprints pertain to a low-dimensional acoustic manifold. This is due to the fact that the intrinsic degrees of freedom (DoF) in acoustic responses are limited to a small number of variables (e.g., room dimensions, source and microphone positions, and refection coefficients). In a fixed environment and microphone constellation, the acoustic fingerprints intrinsically differ only by the source position. Based on this new paradigm we present data-driven algorithms and inference methodologies for source localization and tracking.

The first attempt to address the localization problem using the manifold paradigm was by developing a data-driven and semi-supervised source localization algorithms based on two-microphone measurements. The aim is to accurately recover the inverse mapping between the acoustic fingerprints and their corresponding locations. The first algorithm is based on an interpolation of training positions with weights that are determined based on the diffusion distance between samples. The second algorithm is based on the concept of manifold regularization in a reproducing kernel Hilbert space (RKHS), which extends the standard supervised estimation framework by adding an extra regularization term, imposing a smoothness constraint on possible solutions with respect to a manifold learned in a data-driven manner.

The mapping between the acoustic channel and the source location can be estimated using a Bayesian inference framework, which is analogous to the manifold regularization approach. In the Bayesian formulation, the mapping is modeled as a Gaussian process with a manifold-based prior, which relies on the geometric structure of the manifold. The Bayesian approach and the regularized optimization problem defined in an RKHS both give rise to the same estimators, provided that the same kernel function is used as the covariance function of the Gaussian process and as the reproducing kernel of the RKHS, respectively.

The Bayesian framework facilitates the extension of the single node (microphone pair) setup to an ad hoc network of several microphone pairs. Each node represents a different viewpoint that may be associated with a specific manifold. The information from the different manifolds is merged by defining a multiple manifold Gaussian process, which is obtained by averaging the individual Gaussian processes defined for each node. The resulting algorithm increases the spatial separation and improves the ability to accurately localize the source, outperforming state-of-the-art localization methods in challenging noise and reverberation conditions.

The Bayesian approach also enabled the extension of the static localization method to a dynamic scenario with a moving source. A new data-driven propagation model of the source movement is derived using a Bayesian formulation. The statistical properties of the acoustic fingerprints on the manifold induce a natural propagation model of the source movement that can replace the widely employed random walk or Langevin models. The commonly-used state-space representation of tracking problems, mainly employed by Kalman filtering methods, served as a convenient platform to unify classical and data-driven methods. Two data-modalities with different properties, extracted from the same microphone measurements, were combined under a unified (extended) Kalman filter. The time-difference of arrival (TDOA) readings of the classical regime and the acoustic fingerprints of the data-driven regime are unified in a hybrid algorithm that alternates between the estimates produced by both. The resulting hybrid algorithm demonstrates accurate tracking in adverse acoustic conditions and outperforms competing methods based on only one data modality, namely a TDOA-based or a learning-based approach.

Compared to most existing data-driven localization methods, the presented methods are semi-supervised, i.e., they can be implemented using a flexible amount of training data with only a small set of measurements with calibrated source positions. These methods also extend to distributed array constellations, dynamic scenarios of moving speakers and can be combined with classical localization approaches in a hybrid manner.

1.5 Outline of Monograph

The remainder of the monograph is organized as follows. Some mathematical background on manifold learning methods is given in Section 2. The localization problem is formulated in Section 3, presenting the measured microphone signals, the features extracted from the measurements and the available training information. Section 4 introduces the paradigm of the acoustic manifold and provides supporting simulation results. Based on this paradigm, two manifold-based localization methods using a single node of two microphones are presented in Section 5, based on the diffusion distance as well as optimization with manifold

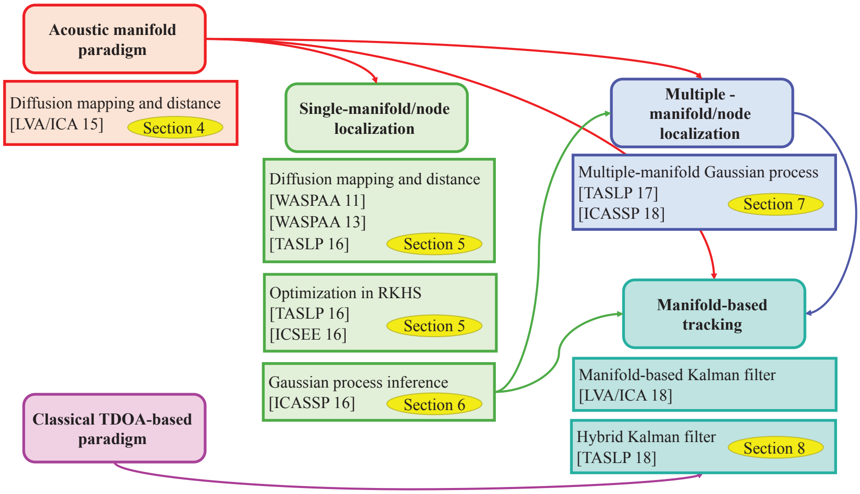

Figure 1.4: Diagram summarizing our latest publications in the field and the relevant sections where they are discussed.

regularization in an RKHS. A Bayesian formulation of the RKHS optimization is discussed in Section 6. Based on this formulation, an extension to multiple-node localization and tracking are presented in Sections 7 and 8, respectively. A diagram summarizing our latest publications in the field and the relevant sections where they are discussed are illustrated in Figure 1.4. A nomenclature listing the different symbols used in this monograph and their meanings is given in Table 1.1.

Table 1.1: Nomenclature.

	Indexes
m	Node (microphone pair) index, $m = \{1, \ldots, M\}$
o	Microphone index in each node, $o = \{1, 2\}$
i	Sample index
t	Continuous/discrete time index, or time-steps of a Markov process
k	Frequency index
	Sizes
n_L	No. of labelled training samples
n_U	No. of unlabelled training samples
n_D	No. of training samples, $n_D = n_L + n_U$
n_T	No. of test samples
n_A	Total no. of samples in both training and test sets, $n_A = n_D + n_T$
D	Dimension of relative transfer functions (RTFs) in the original space
d	Dimension of embedded space, $d \ll D$
	Topological spaces
$\mathcal{M}_m$	The manifold associated with RTFs of the mth node
$\mathcal{H}_k$	Reproducing kernel Hilbert space (RKHS)
	Functions
$\kappa(\cdot, \cdot)$	Standard kernel function measuring similarity between samples
$\tilde{\kappa}(\cdot, \cdot)$	Manifold-based kernel function
$f(\cdot)$	A function mapping between an RTF and one coordinate of source position
$\mathbf{\Phi}_{d,t}(\cdot)$	Diffusion maps of RTFs into an embedding of dimension d and time scale t
	Scalars
$V_o^m(k, \mathbf{p})$	Acoustic transfer function (ATF) relating the source at position $\mathbf{p}$ and the (m, o)th microphone
$H^m(k, \mathbf{p})$	RTF associated with the source at position $\mathbf{p}$ and the mt node, $H^m(k, \mathbf{p}) = V_2^m(k, \mathbf{p})/V_1^m(k, \mathbf{p})$
λ_l	The lth singular-value of transition matrix/graph Laplacian

Continued.

Table 1.1: Continued.

	Vectors
$\mathbf{h}_i^m$	A RTF vector of the mth node and ith sample, $\mathbf{h}_i^m \in \mathbb{R}^D$
$\mathbf{h}_i$	A concatenation of the RTF vectors of all M nodes, $\mathbf{h}_i = \left[[\mathbf{h}_i^1]^T, \ldots, [\mathbf{h}_i^M]^T\right]^T$
$\mathbf{p}_i$	Source position of the ith sample in Cartesian or polar coordinate system
$\boldsymbol{\varphi}_l$	The lth right singular-vector of transition matrix/graph Laplacian
	Matrices
$\mathbf{W}$	Affinity matrix between samples, $W_{ij} = \kappa(\mathbf{h}_i, \mathbf{h}_j)$
$\mathbf{S}$	Degree matrix, $S_{ii} = \sum_{j=1}^n W_{ij}$
$\mathbf{P}$	Transition matrix, $\mathbf{P} = \mathbf{S}^{-1}\mathbf{W}$
$\mathbf{M}$	Graph Laplacian, $\mathbf{M} = \mathbf{S} - \mathbf{W}$
$\mathbf{K}$	Reproducing kernel matrix, $K_{ij} = \kappa(\mathbf{h}_i, \mathbf{h}_j)$
$\boldsymbol{\Sigma}$	Covariance matrix of samples of a Gaussian process, $\boldsymbol{\Sigma}_{ij} = \kappa(\mathbf{h}_i, \mathbf{h}_j)$
$\tilde{\boldsymbol{\Sigma}}$	Manifold-based covariance matrix of samples of a Gaussian process, $\tilde{\boldsymbol{\Sigma}}_{ij} = \tilde{\kappa}(\mathbf{h}_i, \mathbf{h}_j)$

References

[1] Abhayapala, T. D. and H. Bhatta (2003). "Coherent broadband source localization by modal space processing". In: *Proc. of 10th International Conference on Telecommunications*. Vol. 2. French Polynesia: Tahiti. IEEE. 1617–1623.

[2] Adavanne, S., A. Politis, and T. Virtanen (2018). "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network". In: *Proc. of European Signal Processing Conference (EUSIPCO)*. Eurasip. 1462–1466.

[3] Affes, S. and Y. Grenier (1997). "A signal subspace tracking algorithm for microphone array processing of speech". *IEEE Transactions on Speech and Audio Processing*. 5(5): 425–437.

[4] Allen, J. B. and D. A. Berkley (1979). "Image method for efficiently simulating small-room acoustics". *The Journal of the Acoustical Society of America*. 65(4): 943–950.

[5] Aronszajn, N. (1950). "Theory of reproducing kernels". *Transactions of the American Mathematical Society*. 68(3): 337–404.

[6] Beard, M. and S. Arulampalam (2012). "Performance of PHD and CPHD filtering versus JIPDA for bearings-only multi-target tracking". In: *Proc. of the 15th International Conference on Information Fusion (FUSION)*. IEEE. 542–549.

[7] Belkin, M. and P. Niyogi (2003). "Laplacian eigenmaps for dimensionality reduction and data representation". *Neural Computation*. 15(6): 1373–1396.

[8] Belkin, M. and P. Niyogi (2004). "Semi-supervised learning on Riemannian manifolds". *Machine Learning*. 56(1): 209–239.

[9] Belkin, M., P. Niyogi, and V. Sindhwani (2006). "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples". *Journal of Machine Learning Research*. 7(Nov.): 2399–2434.

[10] Belkin, M., P. Niyogi, and V. Sindhwani (2005). "On manifold regularization". In: *Proc. of the 10th International Workshop Artificial Intelligence and Statistics (AISTAT)*. Society for Artificial Intelligence and Statistics.

[11] Benesty, J. (2000). "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization". *The Journal of the Acoustical Society of America*. 107(1): 384–391.

[12] Bérard, P., G. Besson, and S. Gallot (1994). "Embedding Riemannian manifolds by their heat kernel". *Geometric & Functional Analysis GAFA*. 4(4): 373–398.

[13] Berlinet, A. and C. Thomas-Agnan (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.

[14] Bertin, N., S. Kitić, and R. Gribonval (2016). "Joint estimation of sound source location and boundary impedance with physics-driven cosparse regularization". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[15] Blom, H. A. P. and Y. Bar-Shalom (1988). "The interacting multiple model algorithm for systems with Markovian switching coefficients". *IEEE Transactions on Automatic Control*. 33(8): 780–783.

[16] Brandstein, M. S. and H. F. Silverman (1997). "A robust method for speech signal time-delay estimation in reverberant rooms". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. IEEE. 375–378.

[17] Brandstein, M. S., J. E. Adcock, and H. F. Silverman (1997). “A closed-form location estimator for use with room environment microphone arrays”. *IEEE Transactions on Speech and Audio Processing*. 5(1): 45–50.

[18] Breining, C., P. Dreiscitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp (1999). “Acoustic echo control. An application of very-high-order adaptive filters”. *IEEE Signal Processing Magazine*. 16(4): 42–69.

[19] Brendel, A., B. Laufer-Goldshtein, S. Gannot, and W. Kellermann (2019). “Learning-based acoustic source localization using directional spectra”. In: *Proc. of 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE. 276–280.

[20] Brendel, A., B. Laufer-Goldshtein, S. Gannot, R. Talmon, and W. Kellermann (2019). “Localization of an unknown number of speakers in adverse acoustic conditions using reliability information and diarization”. In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 7898–7902.

[21] Bronstein, M. M., J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst (2017). “Geometric deep learning: Going beyond Euclidean data”. *IEEE Signal Processing Magazine*. 34(4): 18–42.

[22] Bross, A., B. Laufer-Goldshtein, and S. Gannot (2020). “Multiple speaker localization using mixture of Gaussian model with manifold-based centroids”. In: *Proc. of 28th European Signal Processing Conference (EUSIPCO)*. Eurasip. 1–4.

[23] Chakrabarty, S. and E. A. P. Habets (2017). “Broadband DOA estimation using convolutional neural networks trained with noise signals”. In: *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 136–140.

[24] Chakrabarty, S. and E. A. P. Habets (2019). “Multi-speaker DOA estimation using deep convolutional networks trained with noise signals”. *IEEE Journal of Selected Topics in Signal Processing*.

[25] Čmejla, J., T. Kounovskỳ, J. Málek, and Z. Koldovskỳ (2018). "Independent vector analysis exploiting pre-learned banks of relative transfer functions for assumed target's positions". In: *Proc. of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Springer. 270–279.

[26] Coifman, R. R. and S. Lafon (2006). "Diffusion maps". *Applied and Computational Harmonic Analysis*. 21(1): 5–30.

[27] Coifman, R. R., S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker (2005). "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps". *Proceedings of the National Academy of Sciences of the United States of America*. 102(21): 7426–7431.

[28] Cox, T. F. and M. A. A. Cox (2000). *Multidimensional Scaling*. Chapman and Hall/CRC.

[29] Dal Degan, N. and C. Prati (1988). "Acoustic noise analysis and speech enhancement techniques for mobile radio applications". *Signal Processing*. 15(1): 43–56.

[30] Datum, M. S., F. Palmieri, and A. Moiseff (1996). "An artificial neural network for sound localization using binaural cues". *The Journal of the Acoustical Society of America*. 100(1): 372–383.

[31] Deleforge, A. and R. Horaud (2012). "2D sound-source localization on the binaural manifold". In: *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 1–6.

[32] Deleforge, A., F. Forbes, and R. Horaud (2013). "Variational EM for binaural sound-source separation and localization". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 76–80.

[33] Deleforge, A., F. Forbes, and R. Horaud (2015). "Acoustic space learning for sound-source separation and localization on binaural manifolds". *International Journal of Neural Systems*. 25(1): 1–19.

[34] DiBiase, J. H., H. F. Silverman, and M. S. Brandstein (2001). "Robust localization in reverberant rooms". In: *Microphone Arrays*. Springer. 157–180.

[35] Dijkstra, E. W. *et al.* (1959). "A note on two problems in connexion with graphs". *Numerische Mathematik*. 1(1): 269–271.

[36] Dmochowski, J., J. Benesty, and S. Affes (2008). "Linearly constrained minimum variance source localization and spectral estimation". *IEEE Transactions on Audio, Speech and Language Processing*. 16(8): 1490–1502.

[37] Do, H., H. F. Silverman, and Y. Yu (2007). "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. IEEE. 121–124.

[38] Doclo, S. and M. Moonen (2003). "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments". *EURASIP Journal on Applied Signal Processing*. 2003: 1110–1124.

[39] Donoho, D. L. and C. Grimes (2003). "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data". *Proceedings of the National Academy of Sciences*. 100(10): 5591–5596.

[40] Dorfan, Y. and S. Gannot (2015). "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources". *IEEE/ACM Transactions on Audio, Speech and Language Processing*. 23(10): 1692–1703.

[41] Dorfan, Y., A. Plinge, G. Hazan, and S. Gannot (2018). "Distributed expectation-maximization algorithm for speaker localization in reverberant environments". *IEEE/ACM Transactions on Audio, Speech and Language Processing*. 26(3): 682–695.

[42] Dvorkind, T. G. and S. Gannot (2005). "Time difference of arrival estimation of speech source in a noisy and reverberant environment". *Signal Processing*. 85(1): 177–204.

[43] Evers, C., A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely (2015). "Bearing-only acoustic tracking of moving speakers for robot audition". In: *Proc. of IEEE International Conference on Digital Signal Processing (DSP)*. IEEE. 1206–1210.

[44] Evers, C., E. A. P. Habets, S. Gannot, and P. A. Naylor (2018). "DoA reliability for distributed acoustic tracking". *IEEE Signal Processing Letters*. 25(9): 1320–1324.

[45] Floyd, R. W. (1962). "Algorithm 97: Shortest path". *Communications of the ACM*. 5(6): 345.

[46] Fourier, B. J. B. J. (1822). *Théorie analytique de la chaleur*. F. Didot.

[47] Gannot, S. and T. G. Dvorkind (2006). "Microphone array speaker localizers using spatial-temporal information". *EURASIP Journal on Advances in Signal Processing*. 2006(1): 1–17.

[48] Gannot, S., D. Burshtein, and E. Weinstein (2001). "Signal enhancement using beamforming and nonstationarity with applications to speech". *IEEE Transactions on Signal Processing*. 49(8): 1614–1626.

[49] Gannot, S., E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov (2017). "A consolidated perspective on multimicrophone speech enhancement and source separation". *IEEE/ACM Transactions on Audio, Speech and Language Processing*. 25(4): 692–730.

[50] Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett (1993). "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1". *STIN*. 93: 27403.

[51] Gebru, I. D., X. Alameda-Pineda, R. Horaud, and F. Forbes (2014). "Audio-visual speaker localization via weighted clustering". In: *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 1–6.

[52] Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *IEE Proceedings F (Radar and Signal Processing)*. Vol. 140. IET. 107–113.

[53] Grondin, F. and F. Michaud (2019). "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations". *Robotics and Autonomous Systems*. 113: 63–80.

[54] Habets, E. A. P. and S. Gannot (2007). "Generating sensor signals in isotropic noise fields". *The Journal of the Acoustical Society of America*. 122(Dec.): 3464–3470.

[55] Habets, E. A. P. (2008). “Noise generators”. URL: https://www.audiolabs-erlangen.de/fau/professor/habets/software/noise-generators.

[56] Habets, E. A. P. (2008). “Room impulse response (RIR) generator”. URL: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

[57] Habets, E. A. P., I. Cohen, and S. Gannot (2008). “Generating nonstationary multisensor signals under a spatial coherence constraint”. *The Journal of the Acoustical Society of America*. 124(5): 2911–2917.

[58] Hadad, E., F. Heese, P. Vary, and S. Gannot (2014). “Multichannel audio database in various acoustic environments”. In: *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE. 313–317.

[59] Haddad, A., D. Kushnir, and R. R. Coifman (2014). “Texture separation via a reference set”. *Applied and Computational Harmonic Analysis*. 36(2): 335–347.

[60] He, W., P. Motlicek, and J.-M. Odobez (2018). “Deep neural networks for multiple speaker detection and localization”. In: *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 74–79.

[61] Hein, M. and J.-Y. Audibert (2005). “Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$”. In: *Proc. of the 22nd International Conference on Machine Learning (ICML)*. ACM. 289–296.

[62] Hornstein, J., M. Lopes, J. Santos-Victor, and F. Lacerda (2006). “Sound localization for humanoid robots-building audio-motor maps based on the HRTF”. In: *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 1170–1176.

[63] Hu, Y., P. N. Samarasinghe, and T. D. Abhayapala (2019). “Sound source localization using relative harmonic coefficients in modal domain”. In: *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 348–352.

[64] Hu, Y., P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot (2020). "Unsupervised multiple source localization using relative harmonic coefficients". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 571–575.

[65] Huang, Y., J. Benesty, and G. W. Elko (2000). "Passive acoustic source localization for video camera steering". In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 2. IEEE 909–912.

[66] Jensen, J. R., M. G. Christensen, and S. H. Jensen (2013). "Nonlinear least squares methods for joint DOA and pitch estimation". *IEEE Transactions on Audio, Speech, and Language Processing*. 21(5): 923–933.

[67] Jensen, J. R., J. K. Nielsen, M. G. Christensen, and S. H. Jensen (2015). "On frequency domain models for TDOA estimation". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 11–15.

[68] Jensen, J. R., J. K. Nielsen, R. Heusdens, and M. G. Christensen (2016). "DOA estimation of audio sources in reverberant environments". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 176–180.

[69] Jolliffe, I. (2011). "Principal component analysis". In: *International Encyclopedia of Statistical Science*. Springer. 1094–1096.

[70] Jones, P. W., M. Maggioni, and R. Schul (2008). "Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels". *Proceedings of the National Academy of Sciences*. 105(6): 1803–1808.

[71] Julier, S. J. and J. K. Uhlmann (1997). "New extension of the Kalman filter to nonlinear systems". In: *Signal Processing, Sensor Fusion, and Target Recognition VI*. Vol. 3068. International Society for Optics and Photonics. 182–194.

[72] Kitić, S., N. Bertin, and R. Gribonval (2014). "Hearing behind walls: Localizing sources in the room next door with cosparsity". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 3087–3091.

[73] Knapp, C. and G. Carter (1976). "The generalized correlation method for estimation of time delay". *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 24(4): 320–327.

[74] Koldovskỳ, Z., J. Málek, and S. Gannot (2015). "Spatial source subtraction based on incomplete measurements of relative transfer function". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 23(8): 1335–1347.

[75] Kounades-Bastian, D., L. Girin, X. Alameda-Pineda, R. Horaud, and S. Gannot (2017). "Exploiting the intermittency of speech for joint separation and diarization". In: *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 41–45.

[76] Kruskal, J. B. (1964). "Nonmetric multidimensional scaling: A numerical method". *Psychometrika*. 29(2): 115–129.

[77] Kumar, L., K. Singhal, and R. M. Hegde (2013). "Robust source localization and tracking using music-group delay spectrum over spherical arrays". In: *IEEE 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE. 304–307.

[78] Kushnir, D., A. Haddad, and R. R. Coifman (2012). "Anisotropic diffusion on sub-manifolds with application to earth structure classification". *Applied and Computational Harmonic Analysis*. 32(2): 280–294.

[79] Lafon, S. and A. B. Lee (2006). "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 28(9): 1393–1403.

[80] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2013). "Relative transfer function modeling for supervised source localization". In: *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 1–4.

[81] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2015). "Study on manifolds of acoustic responses". In: *Proc. of Interntional Conference on Latent Variable Analysis and Signal Seperation (LVA/ICA)*. Springer. 203–210.

[82] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2016). "Manifold-based Bayesian inference for semi-supervised source localization". In: *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6335–6339.

[83] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2016). "A real-life experimental study on semi-supervised source localization based on manifold regularization". In: *Proc. of IEEE International Conference on the Science of Electrical Engineering (ICSEE)*. IEEE. 1–5.

[84] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2016). "Semi-supervised sound source localization based on manifold regularization". *IEEE Transactions on Audio, Speech, and Language Processing*. 24(8): 1393–1407.

[85] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2017). "Semi-supervised source localization on multiple-manifolds with distributed microphones". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 25(7): 1477–1491.

[86] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2017). "Speaker tracking on multiple-manifolds with distributed microphones". In: *Proc. Interntional Conference on Latent Variable Analysis and Signal Seperation (LVA/ICA)*. Springer. 59–67.

[87] Laufer-Goldshtein, B., R. Talmon, I. Cohen, and S. Gannot (2018). "Multi-view source localization based on power ratios". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 71–75.

[88] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2018). "Diarization and separation based on a data-driven simplex". In: *Proc. of 26th European Signal Processing Conference (EUSIPCO)*. Eurasip. 842–846.

[89] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2018). "A hybrid approach for speaker tracking based on TDOA and data-driven models". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 26(4): 725–735.

[90] Laufer-Goldshtein, B., R. Talmon, and S. Gannot (2018). "Source counting and separation based on simplex analysis". *IEEE Transactions on Signal Processing*. 66(24): 6458–6473.

[91] Lederman, R. R. and R. Talmon (2018). “Learning the geometry of common latent variables using alternating-diffusion”. *Applied and Computational Harmonic Analysis*. 44(3): 509–536.

[92] Levy, A., S. Gannot, and E. A. P. Habets (2011). “Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments”. *IEEE Transactions on Audio, Speech, and Language Processing*. 19(6): 1540–1555.

[93] Li, B., T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani (2016). “Neural network adaptive beamforming for robust multichannel speech recognition”. In: *Proc. of Interspeech*. 1976–1980.

[94] Li, X., L. Girin, R. Horaud, S. Gannot, X. Li, L. Girin, R. Horaud, and S. Gannot (2017). “Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization”. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*. 25(10): 1997–2012.

[95] Ma, N., G. Brown, and T. May (2015). “Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions”. In: *Proc. of Interspeech*. Vol. 2015. International Speech Communication Association. 160–164.

[96] Ma, N., T. May, and G. J. Brown (2017). “Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments”. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*. 25(12): 2444–2453.

[97] Ma, W.-K., B.-N. Vo, S. S. Singh, and A. Baddeley (2006). “Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach”. *IEEE Transactions on Signal Processing*. 54(9): 3291–3304.

[98] MacKay, D. J. C. (1998). “Introduction to Gaussian processes”. *NATO ASI Series F Computer and Systems Sciences*. 168: 133–166.

[99] Madhu, N. and R. Martin (2008). “A scalable framework for multiple speaker localization and tracking”. In: *Proc. of International Workshop on Acoustic Signal Enhancment (IWAENC)*.

[100] Mandel, M. I., D. P. Ellis, and T. Jebara (2007). “An EM algorithm for localizing multiple sound sources in reverberant environments”. In: *Proc. of Advances in Neural Information Processing Systems (NIPS).* Curran Associates, Inc. 953–960.

[101] Mandel, M. I., R. J. Weiss, and D. P. W. Ellis (2010). “Model-based expectation-maximization source separation and localization”. *IEEE Transactions on Audio, Speech, and Language Processing*. 18(2): 382–394.

[102] Markovich, S., S. Gannot, and I. Cohen (2009). “Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals”. *IEEE Transactions on Audio, Speech, and Language Processing*. 17(6): 1071–1086.

[103] Markovich-Golan, S., S. Gannot, and I. Cohen (2010). “Subspace tracking of multiple sources and its application to speakers extraction”. In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 201–204.

[104] Markovich-Golan, S., S. Gannot, and W. Kellermann (2016). “Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments”. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 25(2): 320–332.

[105] Markovich-Golan, S., S. Gannot, and W. Kellermann (2018). “Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function”. In: *Proc. of 26th European Signal Processing Conference (EUSIPCO).* Eurasip. 2499–2503.

[106] May, T., S. van de Par, and A. Kohlrausch (2011). “A probabilistic model for robust localization based on a binaural auditory front-end”. *IEEE Transactions on Audio, Speech, and Language Processing*. 19(1): 1–13.

[107] Mercer, J. (1909). “Functions of positive and negative type, and their connection with the theory of integral equations”. *Philosophical Transactions of the Royal Society of London (A).* 209: 415–446.

[108] Minh, H. Q. and V. Sindhwani (2011). "Vector-valued manifold regularization". In: *Proc. of International Conference on Machine Learning (ICML)*. Citeseer. 57–64.

[109] Minh, H. Q., L. Bazzani, and V. Murino (2016). "A unifying framework in vector-valued reproducing kernel Hilbert spaces for manifold regularization and co-regularized multi-view learning". *The Journal of Machine Learning Research*. 17(1): 769–840.

[110] Nadiri, O. and B. Rafaely (2014). "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 22(10): 1494–1505.

[111] Nadler, B., S. Lafon, R. R. Coifman, and I. G. Kevrekidis (2006). "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems". *Applied and Computational Harmonic Analysis*. 21(1): 113–127.

[112] Nadler, B., S. Lafon, I. Kevrekidis, and R. R. Coifman (2006). "Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators". In: *Proc. of Advances in Neural Information Processing Systems (NIPS)*. MIT Press. 955–962.

[113] Nadler, B., S. Lafon, R. Coifman, and I. G. Kevrekidis (2008). "Diffusion maps-a probabilistic interpretation for spectral embedding and clustering algorithms". In: *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer. 238–260.

[114] Nakadai, K., H. G. Okuno, H. Kitano, *et al.* (2002). "Real-time sound source localization and separation for robot audition". In: *Proc. of IEEE International Conference on Spoken Language Processing*. 193–196.

[115] Nyström, E. J. (1929). *Über die praktische Auflösung von linearen Integralgleichungen mit Anwendungen auf Randwertaufgaben der Potentialtheorie*. Akademische Buchhandlung.

[116] Opochinsky, R., B. Laufer-Goldshtein, S. Gannot, and G. Chechik (2019). "Deep ranking-based sound source localization". In: *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 283–287.

[117] Pearson, K. (1901). “On lines and planes of closest fit to systems of points in space”. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 2(11): 559–572.
[118] Perotin, L., R. Serizel, E. Vincent, and A. Guerin (2019). “CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings”. *IEEE Journal of Selected Topics in Signal Processing*. 13(1): 22–33.
[119] Peterson, P. M. (1986). “Simulating the response of multiple microphones to a single acoustic source in a reverberant room”. *The Journal of the Acoustical Society of America*. 80(5): 1527–1529.
[120] Polack, J.-D. (1993). “Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics”. *Applied Acoustics*. 38(2): 235–244.
[121] Rascon, C. and I. Meza (2017). “Localization of sound sources in robotics: A review”. *Robotics and Autonomous Systems*. 96: 184–210.
[122] Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.
[123] Reindl, K., S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann (2013). “Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios”. In: *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 1–4.
[124] Reuven, G., S. Gannot, and I. Cohen (2008). “Dual-source transfer-function generalized sidelobe canceller”. *IEEE Transactions on Audio, Speech, and Language Processing*. 16(4): 711–727.
[125] Roweis, S. T. and L. K. Saul (2000). “Nonlinear dimensionality reduction by locally linear embedding”. *Science*. 290(5500): 2323–2326.
[126] Roy, R. and T. Kailath (1989). “ESPRIT-estimation of signal parameters via rotational invariance techniques”. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 37(7): 984–995.
[127] Rui, Y. and D. Florencio (2004). “Time delay estimation in the presence of correlated noise and reverberation”. In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 2. IEEE. 133–136.

[128] Sainath, T. N., R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals (2015). “Learning the speech front-end with raw waveform CLDNNs”. In: *Proc. of Interspeech*. 1–5.

[129] Sainath, T. N., R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior (2015). “Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms”. In: *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

[130] Sainath, T. N., A. Narayanan, R. J. Weiss, E. Variani, K. W. Wilson, M. Bacchiani, and I. Shafran (2016). “Reducing the computational complexity of multimicrophone acoustic models with integrated feature extraction”. In: *Proc. of Interspeech*. 1971–1975.

[131] Sainath, T. N., R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani (2016). “Factored spatial and spectral multichannel raw waveform CLDNNs”. In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE.

[132] Sainath, T. N., R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. W. Chin, A. Misra, and C. Kim (2017). “Multichannel signal processing with deep neural networks for automatic speech recognition”. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 25(5): 965–979.

[133] Sainath, T. N., R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. W. Chin, A. Misra, and C. Kim (2017). “Raw multichannel processing using deep neural networks”. In: *New Era for Robust Speech Recognition: Exploiting Deep Learning*. Springer.

[134] Sandryhaila, A. and J. M. F. Moura (2014). “Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure”. *IEEE Signal Processing Magazine*. 31(5): 80–90.

[135] Schau, H. C. and A. Z. Robinson (1987). “Passive source localization employing intersecting spherical surfaces from time-of-arrival differences”. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 35(8): 1223–1225.

[136] Scheuing, J. and B. Yang (2008). “Disambiguation of TDOA estimation for multiple sources in reverberant environments”. *IEEE Transactions on Audio, Speech, and Language Processing*. 16(8): 1479–1489.

[137] Schmidt, R. O. (1986). “Multiple emitter location and signal parameter estimation”. *IEEE Transactions on Antennas and Propagation*. 34(3): 276–280.

[138] Schölkopf, B., R. Herbrich, and A. J. Smola (2001). “A generalized representer theorem”. In: *Proc. of the 14th Annual Conference on Computational Learning Theory (COLT)*. Springer. 416–426.

[139] Schwartz, O. and S. Gannot (2013). “Speaker tracking using recursive EM algorithms”. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 22(2): 392–402.

[140] Shalvi, O. and E. Weinstein (1996). “System identification using nonstationary signals”. *IEEE Transactions on Signal Processing*. 44(8): 2055–2063.

[141] Shuman, D. I., S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst (2013). “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains”. *IEEE Signal Processing Magazine*. 30(3): 83–98.

[142] Sindhwani, V., P. Niyogi, and M. Belkin (2005). “Beyond the point cloud: From transductive to semi-supervised learning”. In: *Proc. of the 22nd International Conference on Machine Learning (ICML)*. ACM. 824–831.

[143] Sindhwani, V., W. Chu, and S. S. Keerthi (2007). “Semi-supervised Gaussian process classifiers”. In: *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*. 1059–1064.

[144] Singer, A. and R. R. Coifman (2008). “Non-linear independent component analysis with diffusion maps”. *Applied and Computational Harmonic Analysis*. 25(2): 226–239.

[145] Stéphenne, A. and B. Champagne (1997). "A new cepstral prefiltering technique for estimating time delay under reverberant conditions". *Signal Processing*. 59(3): 253–266.

[146] Takeda, R. and K. Komatani (2016). "Discriminative multiple sound source localization based on deep neural networks using independent location model". In: *Proc. of IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 603–609.

[147] Takeda, R. and K. Komatani (2016). "Sound source localization based on deep neural networks with directional activate function exploiting phase information". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*. 405–409.

[148] Takeda, R. and K. Komatani (2017). "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2217–2221.

[149] Talantzis, F., A. Pnevmatikakis, and A. G. Constantinides (2008). "Audio–visual active speaker tracking in cluttered indoors environments". *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. 38(3): 799–807.

[150] Talmon, R. and R. R. Coifman (2013). "Empirical intrinsic geometry for nonlinear modeling and time series filtering". *Proceedings of the National Academy of Sciences*. 110(31): 12535–12540.

[151] Talmon, R. and S. Gannot (2013). "Relative transfer function identification on manifolds for supervised GSC beamformers". In: *Proc. of 21st European Signal Processing Conference (EUSIPCO)*. Eurasip. 1–5.

[152] Talmon, R., I. Cohen, and S. Gannot (2009). "Relative transfer function identification using convolutive transfer function approximation". *IEEE Transactions on Audio, Speech, and Language Processing*. 17(4): 546–555.

[153] Tenenbaum, J. B., V. De Silva, and J. C. Langford (2000). "A global geometric framework for nonlinear dimensionality reduction". *Science*. 290(5500): 2319–2323.

[154] Teutsch, H. and W. Kellermann (2005). “EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams”. In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 3. IEEE. 89–92.

[155] Thiergart, O., M. Taseska, and E. A. P. Habets (2014). “An informed parametric spatial filter based on instantaneous direction-of-arrival estimates”. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 22(12): 2182–2196.

[156] Tian, Y., Z. Chen, and F. Yin (2015). “Distributed IMM-unscented Kalman filter for speaker tracking in microphone array networks”. *IEEE Transactions on Audio, Speech, and Language Processing*. 23(10): 1637–1647.

[157] Valin, J.-M., F. Michaud, J. Rouat, and D. Létourneau (2003). “Robust sound source localization using a microphone array on a mobile robot”. In: *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vol. 2. IEEE. 1228–1233.

[158] Vermaak, J. and A. Blake (2001). “Nonlinear filtering for speaker tracking in noisy and reverberant environments”. In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 5. IEEE. 3021–3024.

[159] Wabnitz, A., N. Epain, C. Jin, and A. Van Schaik (2010). “Room acoustics simulation for multichannel microphone arrays”. In: *Proc. of the International Symposium on Room Acoustics (ISRA)*. 1–6.

[160] Wakabayashi, M., H. G. Okuno, and M. Kumon (2020). “Multiple sound source position estimation by drone audition based on data association between sound source localization and identification”. *IEEE Robotics and Automation Letters*. 5(2): 782–789.

[161] Wang, Y. and B. Chaib-Draa (2013). “A KNN based Kalman filter Gaussian process regression”. In: *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*. 1771–1777.

[162] Ward, D. B., E. A. Lehmann, and R. C. Williamson (2003). “Particle filtering algorithms for tracking an acoustic source in a reverberant environment”. *IEEE Transactions on Speech and Audio Processing*. 11(6): 826–836.

[163] Weisberg, K. and S. Gannot (2019). "Multiple speaker tracking using coupled HMM in the STFT domain". In: *Proc. of IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE. 286–290.

[164] Weisberg, K., S. Gannot, and O. Schwartz (2019). "An online multiple-speaker DOA tracking using the Cappé-Moulines recursive expectation-maximization algorithm". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 656–660.

[165] Williams, C. K. I. and D. Barber (1998). "Bayesian classification with Gaussian processes". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20(12): 1342–1351.

[166] Woodbury, M. A. (1950). "Inverting modified matrices". *Memorandum Report*. 42: 106.

[167] Wu, X., D. S. Talagala, W. Zhang, and T. D. Abhayapala (2016). "Spatial feature learning for robust binaural sound source localization using a composite feature vector". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6320–6324.

[168] Xiao, X., S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li (2015). "A learning-based approach to direction of arrival estimation in noisy and reverberant environments". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 76–80.

[169] Xiao, X., S. Zhao, T. N. T. Nguyen, D. L. Jones, E. S. Chng, and H. Li (2016). "An expectation-maximization eigenvector clustering approach to direction of arrival estimation of multiple speech sources". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 6330–6334.

[170] Yalta, N., K. Nakadai, and T. Ogata (2017). "Sound source localization using deep learning models". *Journal of Robotics and Mechatronics*. 29(1): 37–48.

[171] Yao, K., J. C. Chen, and R. E. Hudson (2002). "Maximum-likelihood acoustic source localization: Experimental results". In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 3. IEEE. 2949–2952.

[172] Yilmaz, O. and S. Rickard (2004). "Blind separation of speech mixtures via time-frequency masking". *IEEE Transactions on Signal Processing*. 52(7): 1830–1847.

[173] Zelnik-Manor, L. and P. Perona (2005). "Self-tuning spectral clustering". In: *Advances in Neural Information Processing Systems (NIPS)*. 1601–1608.

[174] Zhang, Q., Z. Chen, and F. Yin (2017). "Speaker tracking based on distributed particle filter in distributed microphone networks". *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 47(9): 2433–2443.

[175] Zhong, X. and J. R. Hopgood (2008). "Nonconcurrent multiple speakers tracking based on extended Kalman particle filter". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 293–296.

[176] Zhong, X., A. Mohammadi, A. B. Premkumar, and A. Asif (2015). "A distributed particle filtering approach for multiple acoustic source tracking using an acoustic vector sensor network". *Signal Processing*. 108: 589–603.

[177] Zotkin, D., R. Duraiswami, and L. S. Davis (2001). "Multimodal 3-D tracking and event detection via the particle filter". In: *Proc. of IEEE Workshop on Detection and Recognition of Events in Video*. IEEE. 20–27.

[178] Zotkin, D. N., R. Duraiswami, and L. S. Davis (2002). "Joint audio-visual tracking using particle filters". *EURASIP Journal on Applied Signal Processing*. 2002(1): 1154–1164.