Exposé

# Acoustic Scene Analysis using Distributed Microphone Arrays

Christian Schörkhuber

July 21, 2014

## 1 Introduction

Microphone arrays (MA) have gained a lot of popularity for audio acquisition as they can exploit the spatial diversity of an acoustic scene. They allow for localizing and extracting a given sound source or cancelling out interfering sources [Brandstein and Ward, 2001]. Traditional microphone arrays, however, sample the sound field only locally. This often means that distances between sources and the MA are large, resulting in a low signal-to-noise ratio (SNR).

For some applications this problem can be addressed by placing single microphones randomly with large inter-microphone spacing (we refer to this configuration as *microphone network*). Due to the coarse sampling in space, spatial aliasing prevents the application of traditional array processing algorithms and oftentimes a *best-microphone* approach is adopted.

Distributed microphone arrays (DMA) consist of several,[1] usually rather small, microphone arrays that are randomly placed in the area of interest (in 1 the different array configurations are depicted) . That is, MA processing algorithms can be applied to the individual compact arrays[2] and the acquired data can be jointly processed at the fusion centre.

When distances between the nodes are large or wired connections between the nodes and the fusion centre are infeasible, wireless DMA are preferred.[3] The main challenges in wireless DMA processing arise from the limited bandwidth, clock synchronization issues and limited energy resources [Bertrand, 2011b][Bertrand, 2011a]. As transmission of raw audio data from all sensor nodes to a fusion sensor is frequently impossible, each sensor node performs local processing of its sensor signals and reports only metadata or compressed data to the fusion centre or neighbouring nodes.

[1] Depending on the application the number of deployed MA can range from two to several thousand.

[2] In concordance with the large field of distributed signal processing, we sometimes refer to the individual arrays in a DMA as *sensor nodes* or simply *nodes*.

[3] In literature this topology is usually called wireless acoustic sensor network (WASN).

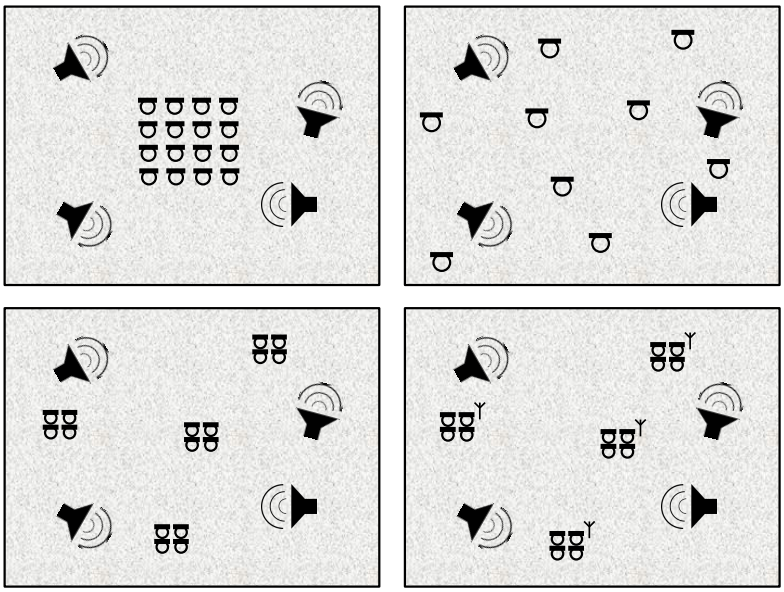

Figure 1: Different array configurations.Upper left: Compact microphone array (MA). Upper right: Single microphone network (MN). Bottom left: Distributed microphone array (DMA). Bottom right: Wireless distributed microphone array (wDMA).

With the Wireless Large-Scale Microphone Array (WiLMA) developed at the *Institute of Electronic Music and Acoustics* (IEM) [Schörkhuber et al., 2014] we are able to develop and evaluate DMA processing algorithms under both synchronized and unsynchronized conditions with different microphone array configurations. We aim to develop new, and enhance existing methods for analysing, transcribing, interpreting, compressing and virtualizing acoustic scenes under real-world conditions including undesired interfering signals and room reverberation.
Applications of robust DMA processing algorithms are numerous, e.g. hands-free telephony, teleconferencing, signal enhancement, dereverberation, acoustic monitoring in domestic, urban, industrial or wildlife scenarios, ambient intelligence, assisted living, room acoustics, intelligent multi microphone recording and virtual acoustics. Consequently, DMA processing and related topics have gained a lot of attention from several research groups in recent years and the amount of literature in this emerging field is substantial. Hence, in section 2 we try to give a coarse (but due to the amount of related sub-topics rather extensive) state-of-the-art report. In section 3 we point out open issues and define our research goals.

# 2 Literature Review

The field of distributed microphone array processing spans numerous sub-problems. In this section we try to give a brief introduction and a coarse state-of-the-art report of topics that are considered relevant for the intended thesis. We discriminate between three classes of problems, namely *system parameter estimation*, *source parameter estimation* and *signal estimation*.

## 2.1 System Parameter Estimation

*System parameters* describe the setting of the scene to be analysed, such as relative or absolute locations of microphones or reflective surfaces.

**Sensor self calibration.** Arguably the most important information to facilitate DMA processing is the spatial relation between microphones. For compact, rigid microphone arrays such as planar or spherical configurations, this information is usually available or can be measured manually. In contrast, for large-scale or ad-hoc[4] microphone arrays, the exact location of each microphone is usually unknown and manual measurement might be impossible or cumbersome at best. Therefore several approaches have been proposed to automate the array calibration process.
In [Sachar et al., 2002] the use of a movable rig with sources in known configuration is suggested, in [Raykar and Duraiswami, 2004] co-located microphones and sources are assumed. The general case where sources at arbitrary unknown locations and unknown emission times are assumed to estimate microphone locations in 2-dimensional space was firstly considered in [Moses et al., 2003]. In [Thrun, 2005] the problem is tackled in 3-dimensional space using the far field assumption, i.e. it is assumed that sources are infinitely far away from the array such that the incoming sound wave hits each sensor at the same incident angle. As this assumption is not valid for many applications, in [Pollefeys and Nister, 2008] a solution to the problem has been proposed that does not rely on this assumption. Based on the results in [Pollefeys and Nister, 2008], in [Gaubitch et al., 2013] an approach is suggested that also accounts for the internal delays in the recording units.
Until very recently, room reverberation was either ignored or considered detrimental for the self-localization problem. In [Dokmanic et al., 2014] early reflections from walls are considered as additional sources that aid the localization procedure. As the location of the source of a reflection leads to additional information about the shape of the room, the approach suggested in [Dokmanic et al., 2014] is reported to locate the microphones' *absolute* positions within the room.
In [Burgess et al., 2013] the authors propose a solution for the self calibration problem when all sources and receivers are assumed to be unsynchronized. They tackle the problem using only unsynchronized time difference of arrival measurements under the far field assumption.

[4] Ad-hoc microphone arrays might consist of devices such as smartphones, tablets, laptops or glasses or may consist of microphones that are deployed for measuring or recording purposes.

**Room geometry inference.** Traditionally reverberation is ignored or considered detrimental in MA processing. On the contrary, recently it has been shown, that knowledge about the shape of the room within which a microphone array operates, can significantly improve the performance of e.g. source localization or source tracking algorithms [Ribeiro et al., 2010a] [Ribeiro et al., 2010b] [Svaizer et al., 2011]. Furthermore this can promote the relative estimation of sound sources and microphones to an absolute one within a given room.
For applications like teleconferencing, auralization, and virtual reality the influence of the room often needs to be compensated for or the illusion of a specific room needs to be created. An accurate model of the early reflections is a prerequisite for the success of these tasks [Lokki and Pulkki, 2002] which requires knowledge about the wall locations.
Room acoustic studies also greatly benefit from augmented RIR where early reflections can be assigned to different walls and reflection orders.
Consequently, geometry inference from measured room impulse responses (RIR) has become an active field of research in recent years. In [Tervo, 2011] [Tervo et al., 2012] the localization and tracing of early reflections has been studied using compact microphone arrays, which is a pre-step to 3D geometry inference. Several approaches have been proposed that require prior information or assume all reflection to be of first order [Antonacci et al., 2010] [Filos et al., 2010] [Filos, 2013]. The authors in [Tervo and Tossavainen, 2012] propose a method to estimate the 3D geometry of a room without any prior assumptions utilizing sparse RIRs obtained by directional room excitation. To our best knowledge, the most general approach to date has been proposed in [Dokmanic et al., 2013]. The authors derive a method to estimate the geometry of convex rooms assuming point sources at unknown locations based on echo labelling by exploiting the properties of Euclidian distance matrices. It is worth mentioning that the authors follow the notion of reproducible research by providing the source code of the proposed algorithm.
In the methods mentioned so far, room impulse responses are measured. A first small step towards geometry inference from continuous signals has been proposed in [Tervo and Korhonen, 2010], where reflective surfaces are estimated from arbitrary excitation signals, e.g. music or speech.
Following the notion in [Tervo and Korhonen, 2010], a very promising approach to geometry inference from arbitrary broadband signals has been proposed in [Mabande et al., 2013]. The authors use spherical microphone arrays to estimate directions of arrival (DOAs) and time differences of arrival (TDOAs) of reflected signals.

## 2.2 Source Parameter Estimation

*Source parameters* provide meta-information about the content of an acoustic scene, such as the number of active sources, their locations and orientations and the corresponding spatial trajectories.

**Source localization.** Estimating the number and locations of acoustic sources has been an active field of research for several decades. Traditionally compact microphone arrays are

used, where the problem reduces to the estimation of directions of arrival (see [Brandstein and Ward, 2001] and references therein). Using DMAs the exact position of an active source in 2 or 3 dimensional space can be estimated.
If fully synchronized (i.e. wired) DMAs are considered, usually time differences of arrival (TDOAs) are utilized with great success. That is, in this case the localization problem boils down to the estimation of TDOAs. The most prominent algorithms for this task are the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [Knapp and Carter, 1976] and the Steered Response Power with Phase Transform (SRP-PHAT) [DiBiase, 2000]. GCC-PHAT and SRP-PHAT have proven to be robust methods for localizing possibly multiple sources in real environments with considerable reverberation[5]. In [Bartsch et al., 2010] different microphone arrays and localization algorithms are evaluated assuming the synchronized case.
A problem that is common to all TDOA based approaches is, that they are valid only in convex rooms where there is a direct line-of-sight between sources and microphones. If the direct sound from a source is missing, an early reflection is considered in the estimation process and the localization must fail. Addressing this problem, a source localization and tracking approach for non-convex rooms has recently been proposed in [Öçal et al., 2014], where the geometry of the room is assumed to be known.
If the nodes of a DMA cannot be synchronized, source locations are estimated by geometrical superposition of multiple DOA estimations from different nodes. The simplest approach is to find intersections of discrete DOA estimations. This method, however, is prone to errors since each node is forced to make a definite estimate about the number of sources and their corresponding DOAs. A more robust approach is to superimpose DOA likelihood functions of all nodes projected into the 2D or 3D space [Tiete et al., 2014]. For complex acoustic scenes both approaches tend to produce *ghost sources* stemming from spurious DOA intersections. This problem can be tackled by iteratively subtracting the most prominent source from the estimation procedure, whereas the problem of a stop criterion seems not to be entirely solved yet.

## 2.3 Signal Estimation

### 2.3.1 Source separation

The extraction of a single audio source from a convolutive mixture is both a very challenging and well researched topic. As it is beyond the scope of this document to give a comprehensive overview about the state-of-the-art of this well matured field of research, the reader is referred to [Vincent et al., 2014] and references therein.
Distributed microphone arrays offer great potential for this task as the disjointness of sources in both the spatial and the time-frequency domain can be exploited. That is, spectral masking can be applied to a reference signal (closest microphone or beamformer output) based on spatial cues [Taseska and Habets, 2014] [Herre et al., 2013] [Taseska and Habets,

[5]SRP-PHAT is more robust to reverberation but less robust at low SNRs.

2013a]. As the use of DMAs extend classical beamformer approaches to the 3-dimensional space, the authors in [Taseska and Habets, 2013b] coined the term *spotforming*.

### 2.3.2 Spatial sound acquisition

By spatial sound acquisition we refer to the problem of finding a compact (parametric) description of a complex acoustic scene to facilitate its transmission, alteration and reproduction. A prominent example that is based on compact microphone arrays is Directional Audio Coding (DirAC) [Pulkki, 2007]. A drawback of this and related methods is, that they are limited to a representation of the sound field with respect to only one point in space, i.e. the location of the array. This location needs to be chosen carefully to both increase the SNR and to capture the spatial image as desired.
The use of DMAs offers new possibilities to capture and describe a sound field with a relatives small number of spatial sampling points. A straight forward parametric description of an acoustic scene is to extract prominent sources along with their spatial location and trajectories augmented with information about the estimated room influence.
An interesting line of research in this context is the extraction of *virtual* microphone signals from an acoustic scene using DMAs [Galdo and Thiergart, 2011] [Thiergart et al., 2013]. The goal is to obtain the signal that a microphone with a given characteristic placed at an arbitrary location in the scene would have recorded.

### 2.3.3 Sound classification and event detection

Detection and classification of acoustic events in larger areas has numerous applications in acoustic surveillance, industrial monitoring, security or assisted living. Pattern recognition algorithms are usually trained on clean data and their performance usually decreases significantly when the desired source is obfuscated by interfering sources, reverberation and noise. If nodes of a DMA are distributed in the area of interest, more robust features can be extracted by applying DMA processing algorithms [Gergen et al., 2014].

# 3 Objectives

In this section we point out some unsolved issues in the field of DMA processing and derive possible research goals. Due to the extent of available literature and the complexity of open issues, the precise subject of our research shall be isolated within the first year of the intended thesis.

## 3.1 Continuous parameter estimation

As discussed in the previous section, distributed microphone array processing has gained a lot of attention among researchers in recent years and solutions to various subtasks have been proposed. Recently, for tasks like self calibration, source localization and source orientation estimation, the influence of room reverberation has been exploited. This seems to be a fruitful line of research, for the free field assumption rarely holds in real world scenarios. As accurate room models are usually the foundation of these methods, inference of the room geometry from acoustic signals is a vital pre-stage for DMA processing. The vast majority of the proposed solutions assume room impulse responses (RIR) to be known *a priori*. While this is a valid assumption for room acoustic applications or fixed DMAs deployed in static environments, RIR are usually unavailable for *ad hoc* networks or DMAs in time-variant environments.
In order to exploit early reflections for DMA processing under non-static conditions, inference of room geometry has to be performed continuously whereas RIRs need to be estimated from unknown arbitrary signals. Since microphone positions need to be known but cannot be assumed to be static either, a joint estimation procedure for both self calibration and geometry inference is desired. The noble goal is to design algorithms that enable DMAs to detect changes of the room geometry or their locations and adapt accordingly without the need for recalibration. The problems that need to be tackled in this endeavour are:

- Estimation of room impulse responses from unknown continuous signals.
- Extending existing algorithms for room geometry inference, self calibration and source localization to the general case of non-convex rooms (including multi-room scenarios and convex rooms populated with non-negligible acoustic obstacles or reflective surfaces)
- Joint estimation of microphone/source positions and room geometry from unknown continuous signals.

## 3.2 Signal extraction

**Reflection exploitation.** Although early reflections are being exploited for source localization, to our best knowledge little research has been done on exploiting reflections for source extraction. This could be of particular interest for directional sources where the accumulated energy of the direct sound might be too low.

**Arbitrary signals.** In the vast majority of methods proposed in the literature, the class of signals under consideration is speech. This stems from the wide range of applications such as teleconferencing, hands-free telecommunication, distant speech recognition and human computer interaction. However, for applications like spatial audio scene description and compression or environmental and industrial monitoring, algorithms need to be developed that cope with a more general class of signals. These signals might not exhibit the same frequently exploited properties like disjointness in standard time-frequency representations.

**Evaluation and metrics.** Furthermore, source separation algorithms designed for speech signals are usually evaluated with dedicated speech quality metrics or word error rates for automatic speech recognition systems. For more general applications other metrics might be more appropriate such as perceived audio quality (e.g. for music signals) or feature robustness for classification tasks.

### 3.3 Array design

Array shapes utilized in DMAs are usually linear, planar, circular or spherical. Spherical arrays have clear advantages in terms of angular range and performance consistency for different impinging angles. However, unobtrusive placing of spherical arrays is quite challenging. Thus, the development of small, wall-mount hemispherical microphone arrays might be considered for real-world applications.

## References


[Antonacci et al., 2010] Antonacci, F., Sarti, A., and Tubaro, S. (2010). Geometric reconstruction of the environment from its response to multiple acoustic emissions. *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2822–2825.

[Bartsch et al., 2010] Bartsch, C., Volgenandt, A., Rohdenburg, T., and Bitzer, J. (2010). Evaluation of different microphone arrays and localization algorithms in the context of ambient assisted living. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4.

[Bertrand, 2011a] Bertrand, A. (2011a). Applications and trends in wireless acoustic sensor networks: A signal processing perspective. *2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, pages 1–6.

[Bertrand, 2011b] Bertrand, A. (2011b). *Signal processing algorithms for wireless acoustic sensor networks.* PhD thesis, KATHOLIEKE UNIVERSITEIT LEUVEN.

[Brandstein and Ward, 2001] Brandstein, M. and Ward, D. (2001). *Microphone arrays: signal processing techniques and applications.* Springer.

[Burgess et al., 2013] Burgess, S., Kuang, Y., and Wendeberg, J. (2013). Minimal Solvers for Unsynchronized TDOA Sensor Network Calibration using Far Field Approximation. In *9th International Symposium on Algorithms and Experiments for Sensor Systems, Wireless Networks and Distributed Robotics (ALGOSENSORS 2013).*

[DiBiase, 2000] DiBiase, J. (2000). *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays.* PhD thesis.

[Dokmanic et al., 2014] Dokmanic, I., Daudet, L., and Vetterli, M. (2014). How to Localize Ten Microphones in a One Fingersnap. In *22nd European Signal Processing Conference.*

[Dokmanic et al., 2013] Dokmanic, I., Parhizkar, R., Walther, A., Lu, Y. M., and Vetterli, M. (2013). Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences of the United States of America*, 110(30):12186–91.

[Filos, 2013] Filos, J. (2013). *Inferring Room Geometries.* PhD thesis, Imperial College London.

[Filos et al., 2010] Filos, J., Habets, E. A. P., and Naylor, P. A. (2010). A two-step approach to blindly infer room geometries. In *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC), Tel Aviv, Israel.*

[Galdo and Thiergart, 2011] Galdo, G. D. and Thiergart, O. (2011). Generating virtual microphone signals using geometrical information gathered by distributed arrays. *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pages 185–190.

[Gaubitch et al., 2013] Gaubitch, N., Kleijn, B., and Heusdens, R. (2013). Auto-localization in ad-hoc microphone arrays. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 106–110.

[Gergen et al., 2014] Gergen, S., Nagathil, A., and Martin, R. (2014). Classification of reverberant audio signals using clustered ad hoc distributed microphones. *Signal Processing*, pages 1–12.

[Herre et al., 2013] Herre, J., Kuech, F., Kallinger, M., Del Galdo, G., and Grill, B. (2013). Apparatus and method for spatially selective sound acquisition by acoustictriangulation, U.S. Patent Application 13/904,857.

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

[Lokki and Pulkki, 2002] Lokki, T. and Pulkki, V. (2002). Evaluation of geometry-based parametric auralization. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*.

[Mabande et al., 2013] Mabande, E., Kowalczyk, K., Sun, H., and Kellermann, W. (2013). Room geometry inference based on spherical microphone array eigenbeam processing. *The Journal of the Acoustical Society of America*, 134(4):2773–89.

[Moses et al., 2003] Moses, R. L., Krishnamurthy, D., and Patterson, R. (2003). A Self-Localization Method for Wireless Sensor Networks. *EURASIP Journal on Applied Signal Processing*, 4(March):348–358.

[Öçal et al., 2014] Öçal, O., Dokmanic, I., and Vetterli, M. (2014). Source Localization and Tracking in Non-Convex Rooms. In *9th International Conference on Acoustics, Speech, and Signal Processing*.

[Pollefeys and Nister, 2008] Pollefeys, M. and Nister, D. (2008). Direct computation of sound and microphone locations from time-difference-of-arrival data. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2445–2448.

[Pulkki, 2007] Pulkki, V. (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, pages 503–516.

[Raykar and Duraiswami, 2004] Raykar, V. C. and Duraiswami, R. (2004). Automatic position calibration of multiple microphones. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*.

[Ribeiro et al., 2010a] Ribeiro, F., Ba, D., Zhang, C., and Florêncio, D. (2010a). Turning enemies into friends: Using reflections to improve sound source localization. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*.

[Ribeiro et al., 2010b] Ribeiro, F., Zhang, C., Florêncio, D., and Ba, D. E. (2010b). Using reverberation to improve range and elevation discrimination for small array sound source localization. *IEEE Transactions on Audio, Speech and Language Processing*, 18(7):1781–1792.

[Sachar et al., 2002] Sachar, J. M., Silverman, H. F., and Patterson, W. R. (2002). Position calibration of large-aperture microphone arrays. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, pages 1797–1800.

[Schörkhuber et al., 2014] Schörkhuber, C., Zaunschirm, M., and Zmölnig, I. (2014). WiLMA-Wireless Largescale Microphone Array. In *Linux Audio Conference 2014*.

[Svaizer et al., 2011] Svaizer, P., Brutti, A., and Omologo, M. (2011). Use of reflected wavefronts for acoustic source localization with a line array. *... Arrays (HSCMA), 2011 Joint ...*, pages 165–169.

[Taseska and Habets, 2013a] Taseska, M. and Habets, E. A. P. (2013a). An online EM algorithm for source extraction using distributed microphone arrays. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, number 1.

[Taseska and Habets, 2013b] Taseska, M. and Habets, E. A. P. (2013b). Spotforming using distributed microphone arrays. In *IEEEWorkshop on Applications of Signal Processing to Audio and Acoustics*.

[Taseska and Habets, 2014] Taseska, M. and Habets, E. A. P. (2014). Informed Spatial Filtering for Sound Extraction using Distributed Microphone Arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(7):1195–1207.

[Tervo, 2011] Tervo, S. (2011). *Localization and tracing of early acoustic reflections*. PhD thesis, Aalto University.

[Tervo and Korhonen, 2010] Tervo, S. and Korhonen, T. (2010). Estimation of reflective surfaces from continuous signals. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 153–156.

[Tervo et al., 2012] Tervo, S., Pätynen, J., and Lokki, T. (2012). Acoustic Reflection Localization from Room Impulse Responses. *Acta Acustica united with Acustica*, 98(3):418–440.

[Tervo and Tossavainen, 2012] Tervo, S. and Tossavainen, T. (2012). 3D room geometry estimation from measured impulse responses. In *Proc. of ICASSP*, pages 513–516.

[Thiergart et al., 2013] Thiergart, O., Galdo, G. D., Taseska, M., and Habets, E. A. P. (2013). Geometry-based Spatial Sound Acquisition Using Distributed Microphone Arrays. *IEEE Transactions on Audio, Speech and Language Processing*, 21(12):2583–2594.

[Thrun, 2005] Thrun, S. (2005). Affine structure from sound. *Advances in Neural Information Processing Systems*, pages 1353–1360.

[Tiete et al., 2014] Tiete, J., Domínguez, F., da Silva, B., Segers, L., Steenhaut, K., and Touhafi, A. (2014). SoundCompass: a distributed MEMS microphone array-based sensor for sound source localization. *Sensors (Basel, Switzerland)*, 14(2):1918–49.

[Vincent et al., 2014] Vincent, E., Bertin, N., Gribonval, R., and Bimbot, F. (2014). From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound. *Signal Processing Magazine, IEEE*, 31(May):107–115.