A Graphical Model Approach to Source Localization in Wireless Sensor Networks

Manish Kushwaha, Xenofon Koutsoukos
Institute for Software Integrated Systems (ISIS)
Department of Electrical Engineering and Computer Science
Vanderbilt University, Nashville, TN 37235, USA
Email: manish.kushwaha@vanderbilt.edu

Abstract

Collaborative localization and discrimination of multiple acoustic sources is an important problem in Wireless Sensor Networks (WSNs). Localization approaches can be categorized as *signal-based* and *feature-based* methods. The signal-based methods are not suitable for collaborative localization in WSNs because they require transmission of raw acoustic data. In feature-based methods, signal features are extracted at each sensor and the localization is done by multisensor fusion of the extracted features. Such methods are suitable for WSNs due to their lower bandwidth requirements. In this paper, we present a feature-based localization and discrimination approach for multiple harmonic acoustic sources in WSNs. The approach uses acoustic beamform and Power Spectral Density (PSD) data from each sensor as the features for multisensor fusion, localization, and discrimination. We use a graphical model to formulate the problem, and employ maximum likelihood and Bayesian estimation for estimating the position of the sources as well as their fundamental and dominant harmonic frequencies. We present simulation and experimental results for source localization and discrimination, to demonstrate our approach. In our simulations, we also relax the source assumptions, specifically the harmonic and omnidirectional source assumptions, and evaluate the effect on localization accuracy. The experimental results are obtained using motes equipped with microphone arrays and an onboard FPGA for computing the beamform and the PSD.

Keywords: Acoustic source localization, Wireless sensor networks, Feature-level fusion, Probabilistic graphical model, Bayesian estimation, Maximum likelihood estimation.

1 Introduction

Acoustic source localization is an important problem in many diverse applications such as military surveillance and reconnaissance, underwater acoustics, seismic remote sensing, and environmental monitoring [1, 2, 3]. Recently, innovative applications such as smart video-conferencing [4], audio-video sensor fusion and target tracking [5, 6] have also been proposed to utilize source localization. Traditional acoustic source localization methods were developed for wired sensor networks [7]. In Wireless Sensor Networks (WSNs), collaborative source localization is used to estimate the positions of multiple sources by fusion of observations from multiple sensors. There are two broad classes of methods for collaborative source localization. The first class of approaches, where the estimation is done by fusion of the raw sampled signals, is called *signal-based* or *signal-level fusion*. The second class of approaches, where signal features are extracted from raw data at each sensor and the estimation is done by fusion of the extracted features, is called *feature-based* or *feature-level fusion*. The signal-level fusion methods are not suited for WSNs because they require transmission of the raw signal, which is costly due to limited bandwidth and power. On the other hand, the feature-level fusion methods are appropriate for WSNs due to their lower bandwidth and power requirements.

In this paper, we present a feature-level fusion approach to collaborative localization and discrimination for multiple harmonic sources in WSNs. Source discrimination involves spatial discrimination where the goal is to separate the sources in space, and frequency discrimination where the goal is to separate the sources in frequency. We use beamforms and Power Spectral Densities (PSDs) as the signal features. The advantage of using the beamform over signal energy is that the beamform captures the angular variation of signal energy, which results in better localization resolution, and hence better spatial source discrimination. Assuming harmonic sources, the use of PSD as another signal feature allows frequency discrimination. Advances in sensor network hardware and FPGA integration has allowed us to implement real-time algorithms for computing such features. Furthermore, the communication bandwidth available in WSNs is sufficient to support wireless transmissions of such features [6].

We use a graphical model to formulate the problem, and employ Maximum Likelihood (ML) and Bayesian estimation for estimating the position of the sources as well as their fundamental and dominant harmonic frequencies. Directed graphical models, which are generalization of Bayesian networks, are directed graphs in which nodes represent random variables, the directed edges represent causality between random variables, and the (lack of) edges capture conditional independence of random variables [8]. We represent the unknown source locations and frequencies as hidden state variables, and the acoustic features as observable variables (or observed data). In directed graphical models, the edges are directed from the hidden state variables to the observed data. Directed graphical models require generative models that describe the observed data in terms of the process that generated them, and the hidden state variables. We present generative models for the beamform and PSD data that describe them in terms of a generative process and the unknown source locations and frequencies.

In our approach, the solution to the collaborative localization and discrimination problem is divided into two steps as source separation and source localization. The idea is to separate the sources in the frequency domain using the PSD data from the sensors, and then use the separated sources for localization and discrimination. We analytically show that source separation is independent of source localization, as long as the sources and the sensors are stationary. We develop a ML estimation method for source separation and use Bayesian estimation for localization. We use Markov Chain Monte Carlo (MCMC) methods, specifically Gibbs sampling and slice sampling [9] for implementing both ML and Bayesian estimation. The advantages of the two step approach instead of joint estimation are twofold. First, estimation in two steps has lower computational complexity than joint estimation. Second, the variances of the likelihood functions for source separation and source localization are significantly different. In context of Monte Carlo methods, joint estimation may cause slower convergence, and require a large number of samples.

We present simulation results for multiple source localization in a grid sensor network. We study three simulation scenarios where (1) we increase the number of sources, (2) we increase the average source SNR of two sources present in the sensing region, and (3) we increase the separation between the two sources. Our results show that as the separation between sources increases, the algorithm is able to achieve higher localization accuracy, comparable to single source localization. We present evaluation of the localization accuracy when the assumptions for the acoustic sources are relaxed; specifically the harmonic and omnidirectional source assumptions. The localization accuracy degrades gracefully when source *harmonicity* is decreased. We also present evaluation of the localization accuracy with PSD data compression. As expected, the localization accuracy improves as more PSD data is available.

Finally, we implement the feature extraction algorithms on an FPGA chip onboard MICAz sensor nodes, and conduct outdoor experiments with real acoustic sources. Outdoor experimental results reinforce the simulation results. For smaller source separations the average localization error remains low but the algorithm is not able to disambiguate the two sources. For larger separations, the localization error is decreased.

The rest of the paper is organized as follows. Related work is present in Section 2. In Section 3, we present the acoustic source model, signal propagation model and feature extraction algorithms. Section 4 describes the graphical model. Sections 5 and 6 describe the source separation and source localization algorithms, respectively. In Section 7, we describe Gibbs sampling and slice sampling. We present results for various simulation scenarios in Section 8, and the outdoor experiment setup and results in Section 9. We conclude in Section 10.

2 Related Work

Signal-based methods for acoustic source localization typically make use of Time Delay of Arrival (TDOA) and Direction of Arrival (DOA). An overview of theoretical aspects of TDOA-based acoustic source localization and beamforming is presented in [10], along with a localization algorithm based on ML estimation. For multiple acoustic sources, an Approximate Maximum Likelihood (AML) algorithm based on alternative projection method is also presented. An empirical study of collaborative acoustic source localization based on an implementation of AML is shown in [3].

Among feature-based methods, Energy-Based Localization (EBL) methods utilize signal energy as the features. Least-squares formulations for EBL have been presented in [11, 12]. A ML formulation with capability for multiple source localization is presented in [13]. They use a multiresolution search algorithm and an expectation-maximization (EM) like iterative algorithm for estimation. We use the beamform instead of the signal energy as the feature. The advantage is that the beamform captures the angular variation of signal energy, which results in better localization resolution, and hence better source discrimination.

The classical approaches to multiple target tracking include data association-based approaches such as Multiple Hypothesis Tracking (MHT) [14] and data association filters [15, 16]. These approaches use a set of exclusive and exhaustive hypotheses either associating measurements with the targets and clutter, called *target-oriented methods*, or associating targets with measurements, called *measurement-oriented methods*. Probabilities are computed for each hypothesis and the most probable hypotheses are used to compute target estimates. The number of hypotheses is combinatorial in the number of targets and measurements, as well as in time.

In data association-based approaches, the *measurements* are noise-corrupted sensor readings related to the state of a target, such as range and/or azimuth from a sensor, etc. The measurements are usually not raw data points, but rather the outputs of signal processing and detection subsystems [16]. The sensor model assumes that each measurement (or detection) corresponds to a single target (i.e. each measurement originated either due to a single target or due to noise). Also, each target generates a single measurement and the measurement due to one target is *not* affected by the presence of other targets. In other words, a measurement for a target would be same regardless the presence of other targets in the scene. Due to these assumptions, the data-association based approaches are not able to model target interaction and mixed measurements, which results in unresolved targets [17]. The problem of mixed measurements and unresolved targets is even more significant in acoustic sensors because the acoustic source signals from multiple targets are additive.

Due to the fact that measurements are not raw data points, but the outputs of signal processing and detection subsystems, the data association-based tracking can be categorized as high-level (or decision-level) fusion. In high-level fusion, *discriminating* information is lost, especially for sensing models where raw data is a mixture of the signals originating from multiple targets. Alternate approaches can utilize low-level (or signal-level) fusion, or medium-level (or feature-level) fusion. Signal-level fusion methods are not suitable for WSNs because they require transmission of the raw signal, which is costly due to limited bandwidth and power. Feature-level fusion methods include signal processing and feature extraction algorithms that *extract* informative features from the raw data, which are communicated to the sensor fusion node. These methods require explicit target interaction models and observation models that describe the generation of features. Feature-level fusion methods are appropriate for WSNs due to their lower bandwidth and power requirements, at the same time, they maintain sufficient *discriminating* information.

More recent approaches to multitarget tracking include joint tracking using Bayesian inference where the quantity of interest is the joint multitarget state, which is the concatenation of individual target states [18]. Joint Bayesian inference has the advantages of providing a recursive solution with arbitrary target dynamic models and observation models. Several approaches based on Bayesian estimation [19, 20] and graphical models [5, 21] have been also proposed. Sequential Monte Carlo (SMC) implementations for Bayesian estimation and multitarget tracking are presented in [22, 23]. A Bayesian approach for tracking the DOA of multiple targets using a passive sensor array is presented in [19]. A Bayesian approach for multiple target detection and tracking, and particle filter-based algorithms are proposed in [20]. A graphical model based approach for audiovisual object tracking that fuses audio and video data from a microphone pair and a camera is presented in [5]. A graphical model formulation for self-localization of sensor networks using a technique called nonparametric belief propagation is presented in [21]. The feature-based localization approach in this paper also uses a graphical model that models multiple target interaction and mixed measurements through generative models for the acoustic features. We use MCMC methods for source separation and localization, and we analytically show that source separation is independent of source localization.

An alternative to Bayesian statistics for multiple target tracking is Finite Set Statistics (FISST). The Bayesian framework treats the states and observations as realizations of random variables. In FISST framework the multitarget state and multiple observations are treated as finite sets. The systematic treatment of multisensor multitarget tracking using random set theory and FISST is presented in [24]. The FISST Bayes multitarget recursion is generally intractable. An approach to approximate the multitarget Bayes recursion by propagating the Probability Hypothesis Density (PHD) of the posterior multitarget state is proposed in [25]. This strategy is similar to the constant gain Kalman filter that propagates the mean of the posterior single-object state. The PHD recursion still involves multiple integrals with no closed forms in general. Several SMC implementations of the PHD filter are proposed in [26, 27]. Like hypothesis-based tracking approaches, the FISST based tracking also assumes one-to-one association between targets and measurements, which results in similar limitation as for the hypothesis-based tracking approaches.

3 Acoustic Source Localization & Discrimination

We consider a WSN of K acoustic sensors in planar field and M far-field stationary acoustic sources coplanar with the sensor network. The objective is to estimate the 2D position of all the sources and discriminate them, both in frequency and space, using the received acoustic signals at the sensors. To localize the acoustic sources, we assume each sensor node is equipped with an array of N_{mic} microphones. The acoustic wave front incident on the microphone array is assumed to be planar for far-field sources. Each sensor receives an acoustic signal that is a combination of source signals. The sensors run signal processing algorithms to compute the beamform and PSD features. In the rest of the section, we describe the source assumptions, source model, signal propagation model and signal processing algorithms for feature extraction.

3.1 Acoustic Source Model

The main assumptions made for the acoustic sources are: (1) omnidirectional and stationary point sources, (2) emitting stationary signals, (3) the source signals are harmonic, and (4) the cross-correlation between two source signals is negligible compared to the signal autocorrelations. Harmonic signals have a fundamental frequency, also called the first harmonic, and other higher-order harmonic frequencies that are multiples of the fundamental frequency. The energy of the signal is contained in these harmonic frequencies only. The harmonic source assumption is satisfied by a wide variety of acoustic sources [28]. In general, any acoustic signal originating due to the vibrations from rotating machinery will have a harmonic structure. The state for the m^{th} acoustic source is given by: (1) the position $\mathbf{x}^{(m)} = \left[x^{(m)}, y^{(m)}\right]^T$, (2) the fundamental frequency $\omega_f^{(m)}$, and (3) the energies in the harmonic frequencies $\psi^{(m)} = \left[\psi_1^{(m)}, \psi_2^{(m)}, \cdots, \psi_H^{(m)}\right]^T$ where H is the number of harmonic frequencies.

In practice, some of the assumptions may not be always true. For example, the engine sound of a vehicle may not be omnidirectional and will be biased toward the side closer to the engine. The physical size of the acoustic source may be too large to be adequately modeled as a point source for sensors very close to the source. In an outdoor environment, strong background noise, including wind gusts, may be encountered during operation. Perhaps the most restrictive assumption is that the source signals are harmonic. In addition to the harmonic components, the engine sound signal may contain other frequency components, which when not accounted for, may cause localization to deteriorate. In Section 8, we present empirical evaluation of localization accuracy when the source assumptions are relaxed, specifically the harmonic and omnidirectional sources assumptions, and it is shown that the localization accuracy degrades gracefully.

3.2 Signal Propagation Model

The intensity of an acoustic signal emitted omni-directionally from a point sound source attenuates at a rate that is inversely proportional to the distance from the source [13]. The discrete signal received at the p^{th} microphone on a particular microphone array is given by

$$r_p[n] = \sum_{m=1}^{M} \frac{d_0}{\parallel \mathbf{x}_p - \mathbf{x}^{(m)} \parallel} s^{(m)}[n - \tau_p^{(m)}] + w_p[n] \tag{1}$$

for samples $n = 1, \cdots, L$, where L is the length of the acoustic signal, M is the number of sources, $w_p[n]$ is white Gaussian measurement noise such that $w_p[n] \sim \mathcal{N}(0, \sigma_w^2)$, $s^{(m)}[n]$ is the intensity of the m^{th} source measured at a reference distance d_0 from that source, $\tau_p^{(m)}$ is the propagation delay of the acoustic signal from the m^{th} source to the p^{th} microphone, and $\mathbf{x}_p$ denote the microphone position. We define the multiplicative term in Equation (1) as the attenuation factor, $\lambda_p^{(m)}$, given by

$$\lambda_p^{(m)} = \frac{d_0}{\parallel \mathbf{x}_p - \mathbf{x}^{(m)} \parallel}.$$

3.3 Acoustic Features

The two acoustic features used for feature-level fusion are beamform and PSD. The details of the feature extraction algorithms are given below. Details related to an FPGA implementation are given in Section 9.

3.3.1 Beamform

Beamforming is a signal processing algorithm for Direction-of-Arrival (DOA) estimation of a signal source using an array of microphone [10]. In a typical delay-and-sum single source beamformer, the 2D sensing region is discretized into directions, or *beams* as $\alpha = i\frac{2\pi}{Q}$, where $i = 0, \cdots, Q-1$ and Q is the number of beams. The beamformer computes the energy of the reconstructed signal at each beam direction. This is achieved by delaying and summing the individual microphone signals. The delayed and summed signal is given by

$$r[n] = \sum_{p=1}^{N_{mic}} r_p[n + t_{pq}(\alpha)] \tag{2}$$

where α is the beam angle, $r_p[\cdot]$ is the received signal at the p^{th} microphone, q is the index of a reference microphone, N_{mic} is the number of microphones, and $t_{pq}(\alpha)$ is the relative time delay for the p^{th} microphone with respect to the reference microphone q, given by

$$t_{pq}(\alpha) = d_{pq} cos(\alpha - \beta_{pq}) f_s / C$$

where d_{pq} and β_{pq} are the distance and angle between the p^{th} and q^{th} microphones, and f_s and C are signal sampling rate and speed of sound, respectively. The beam energy is given by

$$B(\alpha) = \sum_{n=1}^{L} r[n]^2 = \sum_{n=1}^{L} \left[\sum_{p=1}^{N_{mic}} r_p[n + t_{pq}(\alpha)] \right]^2$$

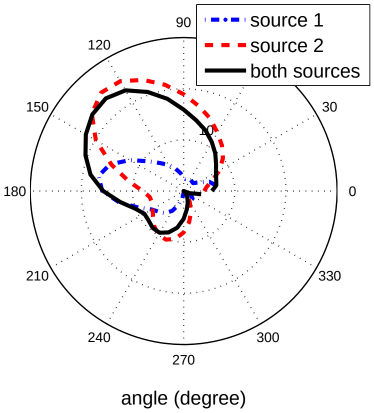
Beam energies are computed for each of the beams, and are collectively called the *beamform*. The beam with maximum energy indicates the DOA of the acoustic source. In case of multiple sources, there might be multiple peaks where the maximum peak would indicate the DOA of the highest energy source. Figure 1(a) shows a beamform for two acoustic sources.

3.3.2 Acoustic PSD

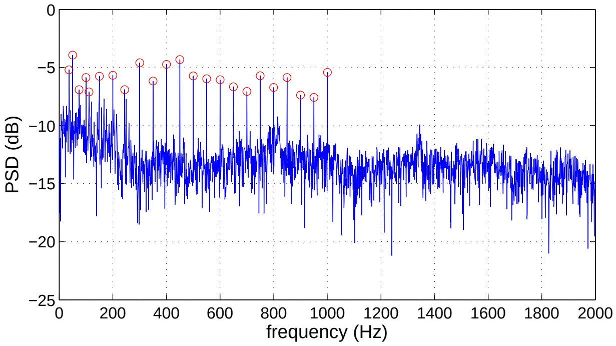
PSD estimation is a signal processing algorithm for estimating the spectrum of the received acoustic signal, which describes how the power of the signal is distributed with frequency [29]. We compute the PSD as the magnitude of the Discrete Fourier Transform (DFT) of the signal

$$P(\omega) = Y(\omega) \cdot \overline{Y(\omega)}$$

where $Y(\omega) = \mathbb{FFT}(r, N_{FFT})$ is the DFT of the signal $r[n]$, N_{FFT} is the length of the transform, and $\overline{Y(\omega)}$ is the complex conjugate of the transform. For real-valued signals, the PSD is real and symmetric; hence we need to store only half of the spectral density. In our implementation, we compress the PSD data for wireless transmission by storing and sending the frequency-power pairs, (ω_j, ψ_j), for N_{PSD} frequencies with the highest power values. Figure 1(b) shows an acoustic PSD estimate for a received signal when two harmonic sources are present.



(a)



(b)

Figure 1: (a) Acoustic Beamforms; The beamforms for single sources clearly show peaks at the source location but the beamform when both sources are present does not show two peaks. (b) Power spectral density (PSD); the highest PSD values are shown as empty circles. The PSD is compactly represented as pairs of the highest PSD values and corresponding frequencies.

4 Graphical Model Overview

Probabilistic graphical models provide a systematic methodology to handle uncertainty and complexity in real systems. They are playing an increasingly important role in the design and analysis of machine learning, bio-informatics, audio processing and image processing algorithms [8, 30]. In a probabilistic graphical model, each node represents a random variable (or a group of random variables), and the edges express probabilistic relationships between these variables. The lack of an edge between nodes represents conditional independence. The graph structure captures the way in which the joint distribution over all the random variables can be factorized into a product of local function defined on nodes and their immediate neighbors (a subset of random variables). Hence, problems involving computation of quantities related to the joint probability model can be profitably cast within a graph theoretic framework. In particular, the underlying structure of the graph is closely related to the underlying computational complexity of an inference problem [8].

The graphical models can be categorized as undirected and directed graphical models. Undirected graphical models capture correlation between random variables, while directed graphical models capture causality between variables. In directed graphical models, also called Bayesian networks, the edges are directed from hidden variables to observed variables. The directed graphical models require *generative models* that describe the observed data in terms of the process that generated them, and the hidden variables.

The model shown in Figure 2 is an example of a directed graphical model. We use this graphical model to formulate the source separation and localization of multiple sources problem. We use plate notation to represent the repetition of the random variables [31]. The plate index, M on the upper plate represents repetition of the hidden variables for M sources, while the index, K on the lower place represents repetition of the observed variables for K sensors. In the figure, the nodes with clear background denote hidden state variables; $\mathbf{x}^{(m)}, \omega_f^{(m)}, \psi^{(m)}$ denote source position, fundamental frequency and harmonic energies for the m^{th} source, respectively. The nodes with shaded backgrounds denote observed variables; B_k and P_k denote the beamform and the PSD received at k^{th} sensor, respectively. Finally, the nodes with dotted outlines denote functions of random variables, or *auxiliary random variables* that capture the functional dependence of the observed variables on the hidden variables. The two auxiliary variables shown in the graphical model are the angle $\theta_k^{(m)}$ and the attenuation factor $\lambda_k^{(m)}$. These variables will be utilized in the generative models for the observed variables.

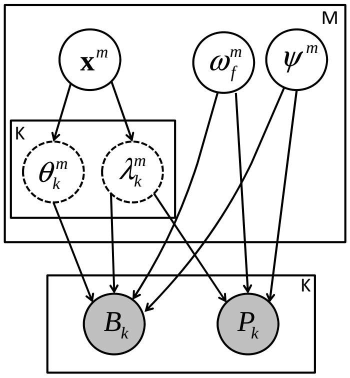


Figure 2: Graphical model.

We perform multiple source localization and discrimination in two steps. First, we use the PSD data only to separate the sources, which in our problem, refers to separating the PSDs of the sources. For harmonic sources, estimation of fundamental frequencies is sufficient for source separation, because all the dominant frequencies in the signal are multiples of the fundamental frequency. An ML estimation method is proposed in the next section for fundamental frequency estimation. It is shown that the ML estimate is independent of the source location, which is intuitive because the dominant frequencies in the source signal are independent of the source location, as long as the source and the sensor are stationary. In the second step, we use the beamform data and the *separated source PSDs* to localize all the sources.

We chose to perform multiple source localization in two steps instead of joint estimation because of the following two reasons. First, estimation in two steps has lower computational complexity than joint estimation. In the context of Monte Carlo methods, let the number of samples needed be exponential in state dimension, $N \propto D^\alpha$, where $D \geq 1$ is the state dimension and $\alpha \geq 1$ is the exponential factor. The joint estimation of two state variables with dimensionality D_1 and D_2 would require $N_{\text{joint}} \propto (D_1 + D_2)^\alpha$ samples, while separate estimation would require $N_{\text{separate}} \propto D_1^\alpha + D_2^\alpha$ samples. For $D_1, D_2, \alpha \geq 1$, one can show that $N_{\text{joint}} \geq N_{\text{separate}}$. Second, the variances of the likelihood functions for source separation and localization are significantly different. In Monte Carlo context, joint estimation may cause slower convergence, and require a large number of samples. Moreover, as mentioned earlier, the ML estimate for source fundamental frequencies is independent of the source locations, further supporting the two step process of source separation and source localization.

5 Source Separation

In this section, we present the first step of our approach, which is source separation and frequency discrimination. We use the PSD data only to separate the sources.We propose an ML estimation method for source separation. In ML estimation, we need the data likelihood function for the PSD data, which requires a generative model. We begin by presenting the generative model and the likelihood function for the PSD data. We also present a result showing that the likelihood function at the ML estimate of harmonic energies is independent of the source positions. Finally, we present the ML estimate for source fundamental frequency.

5.1 Generative Model for PSD Data

For harmonic sources, the PSD can be given by

$$P_s^{(m)}(\omega) = \sum_{h=1}^{H} \psi_h^{(m)} \delta(\omega - h\omega_f^{(m)}) \tag{3}$$

where $m = 1, \cdots, M$ are source indices, ω is the frequency, $\omega_f^{(m)}$ is the fundamental frequency, $\psi_h^{(m)}$ is the energy in the h^{th} harmonic, H is the number of harmonics, and $\delta(\cdot)$ is the Dirac delta function. Using Equation (3), we derive a generative model for the PSD data received at a sensor node. The following proposition states the generative model for the PSD data.[1]

Proposition 1 *For an arbitrary number of acoustic source signals, the power spectral density of the signal received at a sensor is given by*

$$\begin{gathered}\mathbb{P}(\omega) = \sum_{m=1}^{M} \sum_{n=1}^{M} \lambda^{(m)} \lambda^{(n)} \left(P_s^{(m)}(\omega) P_s^{(n)}(\omega) \right)^{\frac{1}{2}} \\ \cos(\Phi^{(m)}(\omega) - \Phi^{(n)}(\omega))\end{gathered} \tag{4}$$

where M is the number of sources, $\lambda^{(m)}$ is the attenuation factor, and $\Phi^{(m)}(\omega)$ is the phase spectral density, which is given by

$$\Phi^{(m)}(\omega) = \phi^{(m)} - \parallel \mathbf{x}^{(m)} - \mathbf{x}_s \parallel \omega / C$$

where $\phi^{(m)}$ is the phase of the source signal, $\mathbf{x}^{(m)}$ and $\mathbf{x}_s$ are the positions of the source and the sensor, respectively.

[1] The proof of this and all other propositions are given in the appendix.

The expression in Equation (4) can be approximated using the following observation. Since we do not maintain the phase of the signal in the source model (see Section 3), we assume all the phases to be normally distributed with equal mean. The expected value of the cosine of the difference of two normally distributed angles is one, i.e. $E[cos(\Phi_i - \Phi_j)] = 1$. Therefore, Equation (4) can be approximated as

$$\mathbb{P}(\omega) \approx \left[\sum_{m=1}^{M} \lambda^{(m)} \left(P^{(m)}(\omega) \right)^{1/2} \right]^2 \tag{5}$$

5.2 Data Likelihood & ML Estimate

Using Equation (5), the negative log-likelihood for PSD data at the kth sensor is defined as

$$\ell_k(\Omega_f, \Psi, \mathbf{X}) = \frac{1}{\sigma_P^2} \int_{\omega} \parallel P_k(\omega) - \mathbb{P}_k(\omega) \parallel^2 d\omega \tag{6}$$

where $P_k(\omega)$ is the observed PSD at the kth sensor, and

$$\begin{aligned} \Omega_f &= \left[\omega_f^{(1)}, \cdots, \omega_f^{(M)} \right]^T \\ \Psi &= \left[\boldsymbol{\psi}^{(1)}, \cdots, \boldsymbol{\psi}^{(M)} \right]^T \\ \boldsymbol{\psi}^{(m)} &= \left[\psi_1^{(m)}, \cdots, \psi_H^{(m)} \right]^T \\ \mathbf{X} &= \left[\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(M)} \right]^T \end{aligned}$$

Assuming a discrete frequency variable, Equation (6) can be rewritten as

$$\ell_k(\Omega_f, \Psi, \mathbf{X}) = \frac{1}{\sigma_P^2} \sum_{\omega_j} \parallel P_k(\omega_j) - \mathbb{P}_k(\omega_j) \parallel^2$$

The likelihood function for the PSD data at the ML estimate of harmonic energies is independent of the source positions, as stated in the following proposition.

Proposition 2 *The likelihood for the PSD data at the ML estimate of harmonic energies is the likelihood that is given by*

$$\ell_k(\Omega_f, \Psi^{ML}, \mathbf{X}) = \ell_k'(\Omega_f, \mathbf{X}) = \sum_{\omega_j \notin \mathbb{H}(\Omega_f)} (P_k(\omega_j))^2 = \ell_k'(\Omega_f) \tag{7}$$

where $\mathbb{H}$ is the set of all harmonic frequencies for all sources

$$\mathbb{H}(\Omega_f) = \bigcup_m \left[\omega_f^{(m)}, 2\omega_f^{(m)}, \cdots \right]^T$$

and, the likelihood is a function of the source fundamental frequencies only, and is independent of the source positions.

Hence, according to Proposition 2, source separation can be performed independent of source localization. The full negative log-likelihood for all sensors, $\ell'(\Omega_f)$ is defined as

$$\ell'(\Omega_f) = \frac{1}{K}\sum_{k=1}^{K} \ell'_k(\Omega_f)$$

Thus, the ML estimation of the fundamental frequencies can be obtained by minimizing $\ell'(\Omega_f)$

$$\hat{\Omega}_f^{ML} = \arg\min_{\Omega_f} \ell'(\Omega_f) = \arg\min_{\Omega_f} \frac{1}{K}\sum_{k=1}^{K} \sum_{\omega_j \notin \mathbb{H}(\Omega_f)} \left(P_k(\omega_j)\right)^2 \tag{8}$$

Note that Equation (8) is *not* an explicit expression for Ω_f since the set $\mathbb{H}$ is a function of Ω_f on the righthand side of this equation. This motivates the use of an iterative method for ML estimation. We use a Monte Carlo method described in Section 7.

6 Source Localization

Source localization is performed by Bayesian estimation in the graphical model shown in Figure 2, and taking the *maximum a-posteriori* (MAP) estimate of the source positions. The posterior, $p(\mathbf{X}|\mathbf{B})$ of the source positions at the ML estimates for source fundamental frequencies and harmonic energies given the beamform data

$$p(\mathbf{X}|\mathbf{B}) \propto \prod_{k=1}^{K} p(B_k|\mathbf{X}, \hat{\Omega}_f^{ML}, \hat{\Psi}^{ML}) p(\mathbf{X})$$

where $p(B_k|\mathbf{X}, \hat{\Omega}_f^{ML}, \hat{\Psi}^{ML})$ is the likelihood function for beamform data, $\mathbf{X}$ represent joint state for all sources, and $\mathbf{B}$ represent the beamforms for all sensors. The likelihood function requires a generative model for the beamform data. In this section, we present the generative model and three intermediate results pertaining to the model. Finally, we present the likelihood and MAP estimation.

6.1 Generative Model for Beamform

We start by developing a generative model for a beamform for a two-microphone array, single-source case (Proposition 3). We will show that the beamform for an arbitrary microphone array (Proposition 4) and an arbitrary number of sources (Proposition 5) can be composed from the simple two-microphone array, single-source case.

Proposition 3 *Consider a microphone pair separated by distance d and the angle between the x-axis and the line joining the microphones is β. For an acoustic source at angle θ and range r with power spectral density $P(\omega)$, the beamform B at the microphone pair is given by*

$$B(\alpha) = 2\lambda^2(R_{ss}(0) + R_{ss}(\kappa_\alpha)) + 2R_\eta(0) \tag{9}$$

where $R_{ss}(\tau) = \mathbb{FFT}^{-1}(P(\omega))$ for $\tau \in [-\infty, +\infty]$ is the autocorrelation of the source signal, $R_{ss}(0)$ is the signal energy, $R_\eta(0)$ is the noise energy, λ is the attenuation factor, and $\kappa_\alpha = d(\cos(\alpha - \beta) - \cos(\theta - \beta))f_s/C$, where $\alpha \in [0, 2\pi]$ is the beam angle, f_s and C are sampling frequency and speed of sound, respectively.

For an arbitrary microphone-array, the generative model can be extended as follows.

Proposition 4 *For an arbitrary microphone-array of N_{mic} microphones, the beamform is expressed in terms of pairwise beamforms as*

$$B(\alpha) = \sum_{(i,j)\in\mathbf{pa}} B_{i,j}(\alpha) - N_{mic}(N_{mic}-2)(R_\eta(0) + \lambda^2 R_{ss}(0)) \tag{10}$$

where **pa** *is the set of all microphone pairs, $R_{ss}(0)$ is the signal energy, $R_\eta(0)$ is the noise energy, λ is the attenuation factor, and $B_{i,j}$ is the beamform for the microphone pair (i,j) (Equation (9)).*

For an arbitrary number of acoustic sources, the generative model is given by the following proposition.

Proposition 5 *For an arbitrary number of uncorrelated acoustic sources M, the beamform is expressed in terms of single source beamforms as*

$$B(\alpha) = \sum_{m=1}^{M} B_m(\alpha) - N_{mic}(M-1)R_\eta(0) \tag{11}$$

where $R_\eta(0)$ is the noise energy and B_m is the beamform for m^{th} acoustic source (Equation (10)).

Finally, the generative model for arbitrary microphone array and arbitrary number of sources can be obtained by substituting Equations (9) and (10) into Equation (11), which gives

$$\mathbb{B}(\alpha) = 2\sum_{m=1}^{M} {\lambda^{(m)}}^2 \sum_{(i,j)\in\mathbf{pa}} R_{ss}^{(m)}(\kappa_\alpha) + N_{mic}\sum_{m=1}^{M} {\lambda^{(m)}}^2 R_{ss}^{(m)}(0) + N_{mic}R_\eta(0) \tag{12}$$

6.2 Data Likelihood & MAP Estimate

Using Equation (12), the negative log-likelihood for beamform data is given as

$$-\ln p(B_k|\mathbf{X}) = \ell_k(\mathbf{X}) = \frac{1}{\sigma_B^2}\sum_{\alpha} \parallel B_k(\alpha) - \mathbb{B}_k(\alpha) \parallel^2$$

The MAP estimate of the source positions is given by

$$\hat{\mathbf{X}}^{MAP} = \arg\max_{\mathbf{X}} p(\mathbf{X}|\mathbf{B}) \tag{13}$$

Since the generative model for beamform in non-linear, an exact method for state estimation in Equation (13) is not possible and we use Monte Carlo method described in the next section for state estimation.

7 Monte Carlo Estimation

Markov Chain Monte Carlo methods are a class of Monte Carlo methods for sampling complex probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. MCMC approaches are so-named because they use the previous sample values to randomly generate the next sample value, generating a Markov chain (as the transition probabilities between sample values are only a function of the most recent sample value). The state of the chain after a large number of steps is then used as a sample from the desired distribution. The quality of the sample improves as a function of the number of steps. The MCMC methods are more efficient, especially for problems with high-dimensional state-space, than sequential Monte Carlo (SMC) methods, also called particle filters [23]. This is due to the fact that the samples in SMC methods are drawn independently, while samples in MCMC are drawn from a Markov chain. If the desired distribution is highly localized in a high-dimensional state space, most of the independent samples drawn by SMC methods would have low probability. On the other hand, after a sufficient number of steps, samples drawn by MCMC methods would, in fact, be from the desired distribution.

The Metropolis-Hastings (MH) algorithm is the earliest of the MCMC method [32]. The MH algorithm generates a Markov chain using a proposal density which depends on the current sample. The proposed samples are accepted as part of Markov chain according to a acceptance-rejection rule. Another popular MCMC method called the Gibbs sampler is very widely applicable to a broad class of Bayesian problems [33]. The Gibbs sampler is a special case of Metropolis-Hastings sampling wherein the proposed sample is always accepted. This results in lesser rejected samples, hence better efficiency. Gibbs sampling, however, requires that all the conditional distributions of the target distribution can be sampled.

In this paper, we use Gibbs sampling for estimation. Gibbs sampling algorithm works on the idea that while the joint probability distribution is too complex to draw samples from directly, the *univariate* conditional distributions – the distribution when all but one of the random variables are assigned fixed values – are easier to sample. We denote the state vector as $X = \left[x^{(1)}, x^{(2)}, \cdots, x^{(D)}\right]^{\mathrm{T}}$, where D is the number of state variables. The joint density $p(X_t|X_{t-1}, Y_t)$ is sampled using Gibbs sampler by sequentially sampling univariate conditional densities given by

$$x_t^{(k,j)} \sim p(x^{(j)}|X_t^{(k,-j)}, X_{t-1}, Y_t) \tag{14}$$

where k is the index of the sample, $j = 1, \cdots, n$ is the index of state variable currently being sampled, and $X_t^{(k,-j)} = \left[x_t^{(k,1)}, \cdots, x_t^{(k,j-1)}, x_t^{(k-1,j+1)}, \cdots, x_t^{(k-1,D)}\right]^{\mathrm{T}}$ is the set of all state variables except $x^{(j)}$. In many cases, the univariate conditional distribution can be arbitrary and the choice of one-dimensional sampling algorithm to sample from the univariate distribution determines the speed and convergence of the Gibbs sampler. We select slice sampling for its robustness in parameters such as step size and applicability toward non-log-concave densities, which is the case in our problem due to multimodal probability distributions [34].

The pseudo code in Algorithm 1 presents the Monte Carlo method for source separation and localization using Gibbs sampling and slice sampling. At time $t = 0$ (line 1), the Gibbs sampling algorithm is initialized with source fundamental frequencies and source locations. For each time $t > 0$, we perform ML estimation for source separation (lines 3–10) and Bayesian estimation for source localization (lines 11–19). The Gibbs sampler draws N samples (line 4 & 12) from the target distribution by sequentially drawing from *univariate* conditional distributions (lines 5–7 & 13–16). Note that slice sampling is used for univariate sampling (lines 6 & 14–15). Notation from Equation (14) is used in lines 6 & 14–15. The state for k^{th} sample is shown in lines 8 & 17. For ML estimation, the sample with minimum negative-log-likelihood is selected as the estimate (line 10). For Bayesian estimation, the sample with maximum *a posteriori* is selected as the state estimate (line 19).

8 Simulation Results

Typically, localization of an acoustic source in WSNs is performed by the sensors that are close to the source because the signal-to-ratio (SNR) is lower for farther sensors. For this reason, we assume

Algorithm 1 Monte Carlo source separation and source localization algorithm

```
At $t = 0$, initialize Gibbs sampler $(\Omega_{f,0}, \mathbf{X}_0)$
for $t > 0$ do
  %%% Source Separation (MC-ML Estimation)
  for $k = 1, \cdots, N$ do
    for $m = 1, \cdots, M$ do
      sample $\omega^{(k,m)}_{f,t} \sim p(\omega^{(m)}_f | \Omega^{(k,-m)}_{f,t}, \Omega_{f,t-1}, P_t)$
    end for
    $\Omega^{(k)}_{f,t} = \left[\omega^{(k,1)}_{f,t}, \cdots, \omega^{(k,M)}_{f,t}\right]$
  end for
  ML estimate, $\hat{\Omega}^{ML}_{f,t} = \arg\min_{\Omega^{(k)}_{f,t}} \ell'(\Omega^{(k)}_{f,t})$
  %%% Source Localization (MC Bayesian Estimation)
  for $k = 1, \cdots, N$ do
    for $m = 1, \cdots, M$ do
      sample $x^{(k,m)}_t \sim p(x^{(m)} | \mathbf{X}^{(k,-m)}_t, \mathbf{X}_{t-1}, B_t, \hat{\Omega}^{ML}_{f,t})$
      sample $y^{(k,m)}_t \sim p(y^{(m)} | \mathbf{X}^{(k,-m)}_t, \mathbf{X}_{t-1}, B_t, \hat{\Omega}^{ML}_{f,t})$
    end for
    $\mathbf{X}^{(k)}_t = \left[x^{(k,1)}_t, y^{(k,1)}_t, \cdots, x^{(k,M)}_t, y^{(k,M)}_t\right]$
  end for
  MAP estimate, $\hat{\mathbf{X}}^{MAP}_t = \arg\max_{\mathbf{X}^{(k)}_t} p(\mathbf{X}^{(k)}_t | \mathbf{X}_{t-1}, \mathbf{B})$
end for
```

that even in a large sensor network, a source will be surrounded by a small number of sensors that will participate in the localization of that source.

8.1 Setup and Parameters

In simulations, we consider a sensor network of 4 acoustic sensors arranged in a grid of size $10m \times 5m$, wherein each sensor can detect all the sources. We simulate the sources according to the acoustic source model in Section 3, simulate the data according to the feature extraction algorithms in Section 3, and finally compare the output of source localization against the ground truth. The performance of the approach is measured in terms of the localization error, which is defined as the root mean square (RMS) position error averaged over all the sources

$$E = \frac{1}{M} \sum_{m=1}^{M} ||\mathbf{x}^{(m)} - \tilde{\mathbf{x}}^{(m)}||$$

where M is the number of source, and $\mathbf{x}^{(m)}$ and $\tilde{\mathbf{x}}^{(m)}$ are the estimated and the ground truth positions for the m^{th} source, respectively. Table 1 shows the parameters used in the algorithm.

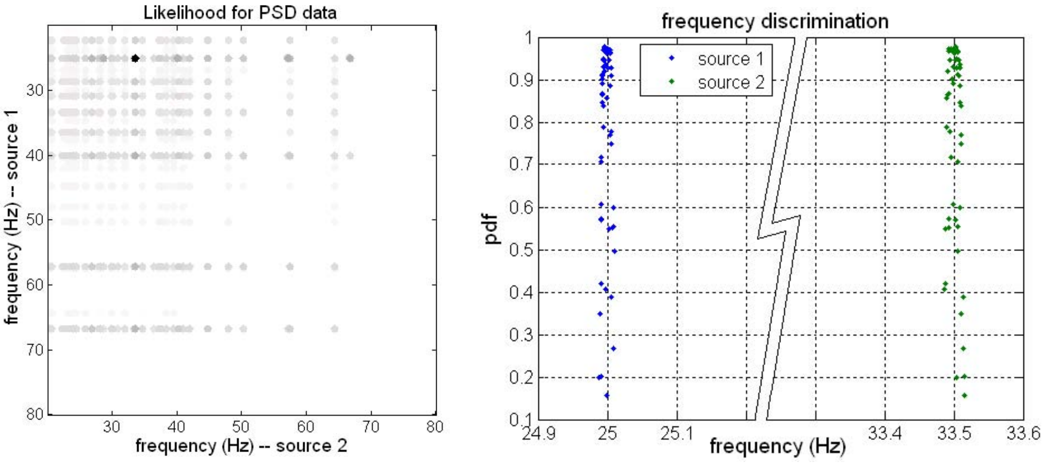
8.2 Frequency Discrimination

Figure 3(a) shows the PSD data likelihood for two sources. The data likelihood is highly multimodal but localized. The data likelihood near the true fundamental frequency values is higher than the likelihood for other frequencies. Figure 3(b) shows the localized nature of the PSD data likelihood. The data likelihood is centered around the true fundamental frequency with a standard deviation

Table 1: Parameters used in simulations

Sampling frequency (f_s)	100kHz
Speed of sound (C)	350 m/sec
Audio data length (time)	1 sec
Maximum harmonic frequency (ω_{max})	1000Hz
SNR (dB)	25
Number of beams	36
Size of Fourier transform (N_{FFT})	4000
Number of Gibbs samples	40

of less than 0.01 Hz. Due to this localized nature of the data likelihood, we can discriminate fundamental frequencies as close as 0.1 Hz.



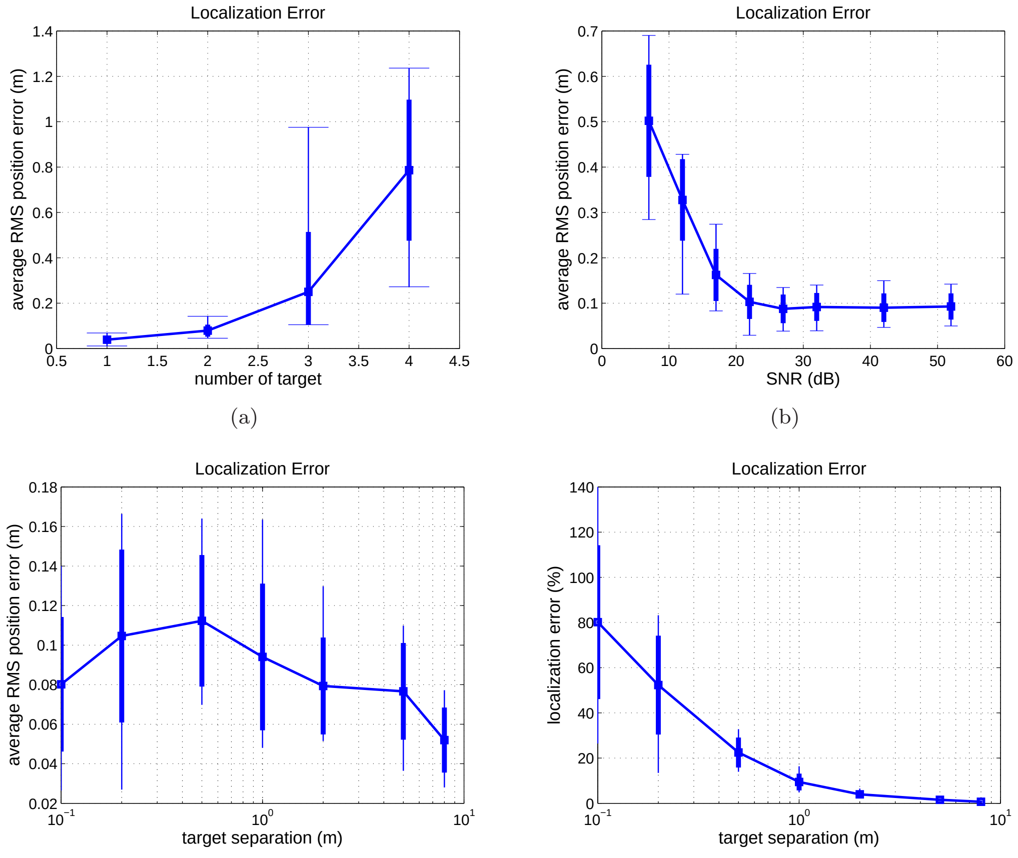
(a) (b)

Figure 3: (a) PSD data likelihood, (b) Frequency discrimination.

8.3 Localization and Spatial Discrimination

We study three simulation scenarios. In the first scenario, we increase the number of sources present in the sensing region gradually to see the effect on localization accuracy. In the second scenario, we increase the average source SNR of two sources present in the sensing region. In the third scenario, we increase the separation between two sources present in the sensing region to evaluate spatial discrimination.

Figure 4(a) shows the localization error for the first scenario when the number of sources is increased from 1 to 4. The localization error increases approximately exponentially with the number of sources. Figure 4(b) shows the average localization error for the second scenario when source SNR for the two sources is increased from 7dB to 52dB. As expected, the localization error decreases with increasing SNR and remains approximately constant above 25dB. Figures 4(c) and 4(d) show the localization error for the third scenario when the source separation between the two sources is increased from 0.1m to 8m. For small source separations (0.1m and 0.2m), the localization error is



(a)

(b)

(c)

(d)

Figure 4: Localization error with (a) Source density, (b) Source SNR, and (c) Source separation. (d) Localization error as a percentage of source separation.

of the same order as the separation. This indicates that the two sources cannot be disambiguated at such separation. For higher source separation (above $0.5m$), the localization error is a small fraction of the separation distance. This indicates that the two sources are successfully localized and discriminated. In fact, for larger source separation (above $5m$), the average localization error for the two sources is the same as that of the single sources.

8.4 Relaxation of Source Assumptions

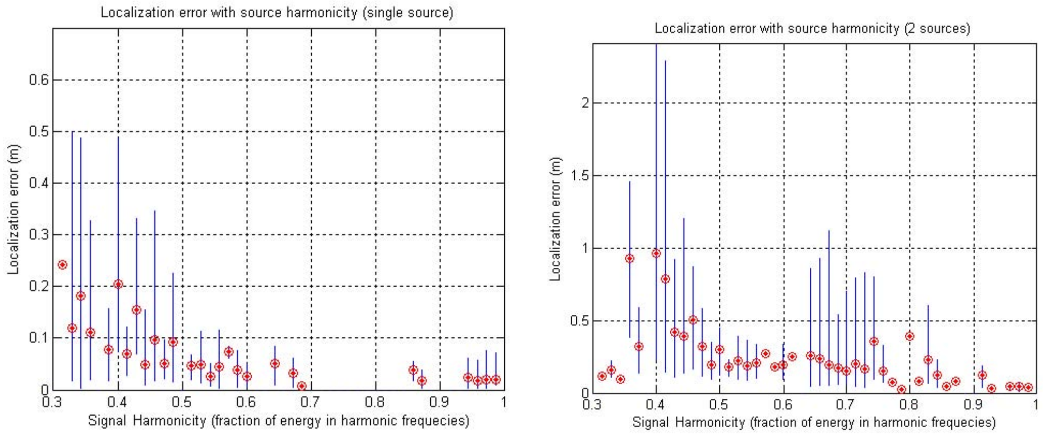
We evaluate the localization accuracy when the assumptions for the acoustic sources are relaxed; specifically the harmonic and omnidirectional sources assumptions.

8.4.1 Harmonicity

Harmonicity of a signal represents the degree of acoustic periodicity in the signal. In this analysis, we define signal harmonicity as the ratio of signal energy in the harmonic frequencies over the total signal energy

$$\hbar = \frac{\sum_{\omega \in \mathbb{H}} P(\omega)}{\sum_{\omega} P(\omega)} \tag{15}$$

Acoustic signals originating from rotating machinery will have high harmonicity close to unity, while signals due wind or a white noise source will have low harmonicity. The localization accuracy of our approach degrades gracefully when signal harmonicity is decreased. Figure 5(a) and 5(b) show the localization error for single source and two sources, respectively, with signal harmonicity. The localization error decreases as the signal harmonicity is increased. Hence, the signal harmonicity computed using Equation (15) can be also used as an indicator of confidence in the localization result.



(a) (b)

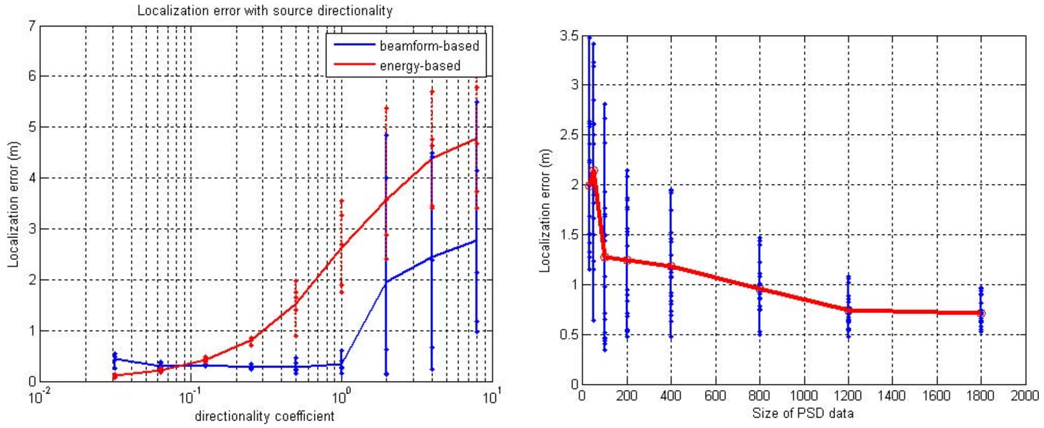
Figure 5: Localization error with signal harmonicity for (a) Single Source, (b) Two Sources.

8.4.2 Directivity

Directivity of an acoustic source is a measure of its directional characteristics. Directivity indicates how much signal energy is directed toward a specific area compared to the total signal energy being transmitted by the source. In this analysis, we express 2D source directivity in terms of a *directionality coefficient* that governs the attenuation of the signal energy with the angle. The directional signal attenuation is given by

$$\lambda_{\phi} = \left(\frac{2 + \cos \phi}{3} \right)^{\beta}$$

where β is the directionality coefficient and ϕ is the direction. The value of directionality coefficient is zero ($\beta = 0$) for omnidirectional sources (i.e. the signal attenuates uniformly in all directions). Figure 6(a) shows the localization error for two sources when the directionality of the sources is increased. The localization error for beamform-based localization is not affected for directionality coefficient as high as 1.0, after which the error increases rapidly. As compared to energy-based localization, beamform-based localization is able to cope with higher directionality. It is expected that if we increase the number of sensors in beamform-based localization, the effect of source directionality can be further reduced.



(a)

(b)

Figure 6: Localization error with (a) Source directionality, (b) Amount of PSD data available.

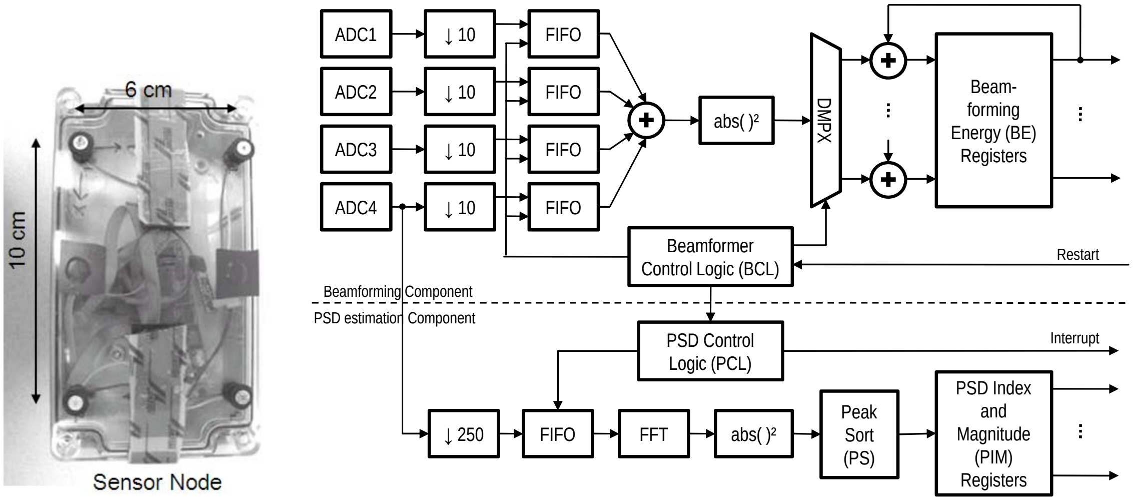
8.5 PSD Data Compression

We also present empirical analysis for the effect of PSD data compression to the localization accuracy by increasing the N_{PSD} parameter presented in Section 3.3.2. As expected, the localization accuracy improves when more PSD data is available. Figure 6(b) shows the localization error for two sources as a function of size of PSD data available to the fusion algorithm. As expected, the localization error decreases as more PSD data is made available to the base station. The accuracy is poor and degrades rapidly for smaller PSD data size.

9 Outdoor Experiments

We implemented the beamforming and PSD estimation algorithms described in Section 3 on a Xilinx XC3S1000 FPGA based MICAz sensor motes (see Figure 7(a)). The outline of the overall design is as follows. First, the FPGA collects samples of the signal received at the microphones in FIFO buffers and performs beamforming and PSD estimation. Then the FPGA stores the results, i.e. beamforming energies and N_{PSD} highest values of the PSD, in registers easily accessible from the MICAz mote. Once these registers are read the mote transmits their value to the base station where further processing and sensor fusion takes place. The block diagram of the FPGA design is shown in Figure 7(b), where the upper half (above the dashed line) represents the beamforming- and the bottom half shows the PSD estimation component. Beamforming component utilizes 166 msec of audio data each cycle, while the PSD estimation component utilizes 1 sec of data with 75% overlap. The angular resolution of beamforming is 10 degrees while frequency resolution of PSD estimation is 1 Hz. The PSD estimation component returns 30 PSD values.

Beamforming Component. The entire FPGA application runs at 20 MHz and the Analog to Digital Converters (ADCs) are sampling at 1 MSPS. The beamforming component uses all four microphone channels. The 8-bit ADC values are first downsampled by a factor of 10 and stored in a FIFO buffer. FIFO buffers are realized as circular buffers using 2-Kbyte embedded block RAMs that store sample values from last 20 ms. Read access to the FIFO buffers and write access to



(a) (b)

Figure 7: (a) Acoustic sensor node with 4 microphone. (b) Block diagram of Beamforming- and PSD estimation components realized on FPGA.

the Beamforming Energy (BE) registers are controlled by the Beamformer Control Logic (BCL). Elements in the FIFOs are selected for read by a simple state machine that resides in the BCL. The state machine iterates through 36 states each corresponding to a specific direction and accesses the FIFO elements based on a look-up-table that contains the delay information for the different angles. The accessed values of the four FIFOs are summed, squared and accumulated in the corresponding BE register in one FPGA clock cycle for each direction. Once the energies for all the 36 directions have been calculated, BCL waits for the next sample to arrive and starts this procedure over. This procedure is repeated 2^{14} times which takes approximately 160 ms and ensures that the 32-bit BE registers do not overflow. After the last cycle, a trigger signal is generated for the PSD estimation component and the beamforming component is halted.

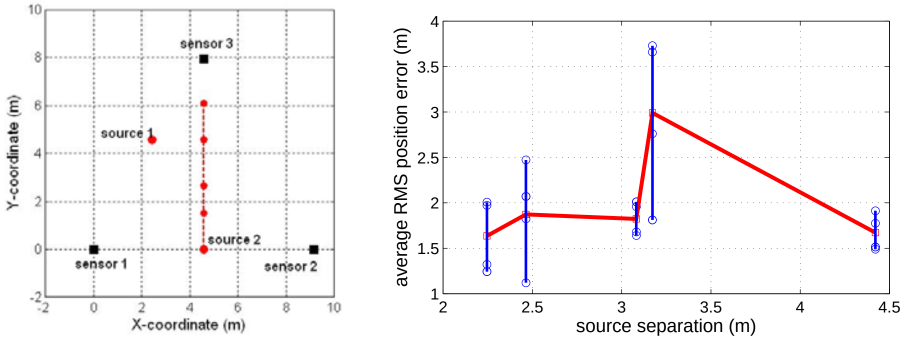
PSD Estimation Component. The PSD estimation component uses the samples from only one channel. The signal is first re-sampled at 4 kHz. For the decimation process, a 600-tap poly-phase filter with 2 kHz cut-off frequency is used. The decimated 8-bit samples are then fed into a 4-Kbyte FIFO buffer which allows to store up to 1 s sample history. In response to a trigger event received from the beamforming component the FIFO content is loaded into the 4096 point FFT module. Only the magnitude information of the first 2048 FFT results is forwarded to the Peak Sort (PS) module. The PS module then selects and stores 30 elements with largest magnitudes and their corresponding indices. After all the FFT samples are processed an interrupt signal is generated for the MICAz mote, which in return reads all the register values through an I2C bus, transmits the values through the radio and restarts the beamforming component. The PSD estimation time is negligible compared to the I2C and radio transfers, which take approximately 90 ms together. The beamforming process in conjunction with the PSD estimation and the data transfers takes roughly 250 ms, which results in an approximately 4 Hz update rate at the base station.

Resource Utilization. The resources used by the beamforming- and PSD estimation components and the overall design (including all driver modules) are shown in Table 2. The utilization rates of the different FPGA components compared to available resources found in the Xilinx XC3S1000 FPGA prove the feasibility of this application, but also points out its limitations. Since 75% of block RAMs are already used, memory resources are likely to become a bottleneck when using larger data sets (FIFO sizes).

	Beamforming	**PSD Estimation**	**Overall Design**
Flip Flops	1,258 (8%)	2,018 (13%)	3,843 (8%)
4 input LUTs	1,398 (9%)	2,578 (16%)	6,860 (44%)
Block RAMs	8 (33%)	9 (38%)	18 (75%)
Hardware multipliers	1 (4%)	6 (25%)	7 (29%)

Table 2: FPGA resource utilization of the Beamforming-, PSD Estimation modules and the overall design in absolute numbers and relative to the available resources found in the Xilinx XC3S1000 FPGA.

Experiments using an Outdoot Deployment We deployed a sensor network of 3 MICAz-based acoustic sensor nodes in an equilateral triangle of side length 9.144m (15ft). Figure 8(a) shows the experimental setup and the location of the sources. We collected the sensor data and ran the algorithm offline. Figure 8(b) shows the localization error with source separation. The results follow the similar trend as that in Figure 4(c). For smaller source separations, the average error remains low but the algorithm is not able to disambiguate the two sources. For larger separations, the localization error decreases.



(a)

(b)

Figure 8: (a) Outdoor experimental setup. Source 1 is kept at the same location while source 2 is placed at different locations. (b) Localization error with source separation.

10 Conclusion

In this paper, we proposed a feature-based fusion method for localization and discrimination of multiple acoustic sources in WSNs. Our approach fused beamforms and PSD data from each sensor. The approach utilized a graphical model for estimating the source positions and the fundamental frequencies. We subdivided the problem into source separation and source localization. We showed in simulation and outdoor experiments that the approach can discriminate multiple sources using the simple features collected from the resource-constrained sensor nodes. As part of an ongoing work, we are working on target dynamics models to extend the approach for multiple source tracking. In the future, the use of graphical models will allow us to extend the approach to multimodal sensors.

References

[1] C. Savarese, J. Rabaey, and J. Beutel, "Location in distributed ad-hoc wireless sensor networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001*, vol. 4, 2001, pp. 2037–2040.

[2] D. Estrin, D. Culler, K. Pister, and G. Sukhatme, "Connecting the physical world with pervasive networks," in *IEEE Pervasive Computing*, vol. 1, no. 1, Jan-Mar 2002, pp. 59–69.

[3] A. M. Ali, K. Yao, T. C. Collier, C. E. Taylor, D. T. Blumstein, and L. Girod, "An empirical study of collaborative acoustic source localization," in *IPSN '07: Proceedings of the 6th international conference on Information processing in sensor networks*, 2007.

[4] B. H. Yoshimi and G. S. Pingali, "A multimodal speaker detection and tracking system for teleconferencing," in *ACM Multimedia '02*, 2002.

[5] M. J. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, 2003.

[6] M. Kushwaha, I. Amundson, P. Volgyesi, P. Ahammad, G. Simon, X. Koutsoukos, A. Ledeczi, and S. Sastry, "Multi-modal target tracking using heterogeneous sensor networks," in *In International Conference on Computer Communications and Networks (ICCCN 2008)*, 2008.

[7] S. S. Haykin and J. H. Justice, *Array Signal Processing*. Prentice-Hall, Inc., 1985.

[8] M. I. Jordan, *Learning in Graphical Models*. The MIT Press, 1998.

[9] C. P. Robert and G. Casella, *Monte Carlo statistical method*. (Springer Texts in Statistics) Springer-Verlag, 2004.

[10] J. C. Chen, K. Yao, and R. E. Hudson, "Acoustic source localization and beamforming: theory and practice," in *EURASIP Journal on Applied Signal Processing*, April 2003.

[11] D. Li and Y. H. Hu, "Energy-based collaborative source localization using acoustic microsensor array," in *EURASIP Journal on Applied Signal Processing*, vol. 2003, 2003.

[12] C. Meesookho, U. Mitra, and S. Narayanan, "On energy-based acoustic source localization for sensor networks," in *IEEE Transactions On Signal Processing*, vol. 56, 2008.

[13] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple source localization using acoustic energy measurements with wireless sensor networks," in *IEEE Transactions On Signal Processing*, vol. 53, 2005.

[14] D. Reid, "An algorithm for tracking multiple targets," vol. 24, no. 6, pp. 843–854, December 1979.

[15] S. S. Blackman, *Multiple-Target Tracking with Radar Applications*. Artech House, 1986.

[16] Y. Bar-Shalom and T. Fortmann, *Tracking and data association.* Academic Press Professional, Inc., 1988.

[17] R. Mahler, “PHD filters for nonstandard targets, II: Unresolved targets,” in *12th International Conference on Information Fusion*, 2009, pp. 922–929.

[18] L. D. Stone, C. A. Barlow, and T. L. Corwin, *Bayesian Multiple Target Tracking.* Artech House, 1999.

[19] M. Orton and W. Fitzgerald, “A bayesian approach to tracking multiple targets using sensor arrays and particle filters,” in *IEEE Transactions on Signal Processing*, vol. 50, 2002.

[20] M. R. Morelande, C. M. Kreucher, and K. Kastella, “A bayesian approach to multiple target detection and tracking,” in *IEEE Transactions on Signal Processing*, vol. 55, 2007.

[21] A. T. Ihler, J. W. Fisher, R. L. Moses, and A. S. Willsky, “Nonparametric belief propagation for self-localization of sensor networks,” in *IEEE Journal on Selected Areas in Communications*, vol. 23, 2005.

[22] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, 2001.

[23] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking,” in *IEEE Transactions on Signal Processing*, vol. 50, 2002.

[24] R. Mahler, *An Introduction to Multisource-Multitarget Statistics and Applications.* Lockheed Martin Technical Monograph, 2000.

[25] ——, “Multi-target bayes filtering via first-order multi-target moments,” in *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, 2003.

[26] B. N. Vo, S. Singh, and A. Doucet, “Sequential monte carlo implementation of the phd filter for multi-target tracking,” in *In Proceedings of the Sixth International Conference on Information Fusion*, 2003.

[27] T. Zajic, R. Ravichandran, R. Mahler, R. Mehra, and M. Noviskey, “Joint tracking and identification with robustness against unmodeled targets,” in *Signal Processing, Sensor Fusion and Target Recognition XII*, vol. 5096, 2003.

[28] C. Serviere and P. Fabry, “Blind source separation of noisy harmonic signals for rotating machine diagnosis,” in *Journal of Sound and Vibration*, vol. 272, no. 1-2, 2004.

[29] B. Porat, *Digital processing of random signals: theory and methods.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1994.

[30] F. V. Jensen, *Bayesian Networks and Decision Graphs.* Springer, 2001.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” in *Journal of Machine Learning Research*, vol. 3. MIT Press, 2003, pp. 993–1022.

[32] W. K. Hastings, “Monte carlo sampling methods using markov chains and their applications,” in *Biometrika*, vol. 57, no. 1, 1970, pp. 97–109.

[33] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, 1984, pp. 721–741.

[34] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 2002.

A Appendix

A.1 Proof of Proposition 1

Proof. Consider M sources emitting source signals $s_m[n]$, for $m = 1, 2, \cdots, M$. Using Equation (1), the received signal at a microphone is given by

$$y[n] = \sum_{m=1}^{M} \lambda_m s_m[n - \tau_m] + w[n]$$

where τ_m is the propagation delay, and λ_m is the attenuation factor. Taking FFT of the received signal, we have

$$\begin{aligned} Y(\omega) &= \mathbb{FFT}\big(y[n]\big) \\ &= \mathbb{FFT}\left(\sum_{m=1}^{M} \lambda_m s_m[n - \tau_m] + w[n]\right) \\ &= \sum_{m=1}^{M} \lambda_m \mathbb{FFT}\big(s_m[n - \tau_m]\big) + \mathbb{FFT}\big(w[n]\big) \\ &= \sum_{m=1}^{M} \lambda_m S_m(\omega) + W(\omega) \end{aligned}$$

where $S_m(\omega)$ is the Fourier transform of m^{th} source signal, and $W(\omega)$ is the Fourier transform of noise. The power spectral density (PSD) of a signal is given by

$$\begin{aligned} P(\omega) &= Y(\omega) \cdot \overline{Y(\omega)} \\ &= \left(\sum_m \lambda_m S_m(\omega) + W(\omega)\right) \cdot \overline{\left(\sum_m \lambda_m S_m(\omega) + W(\omega)\right)} \\ &= \left(\sum_m \lambda_m S_m(\omega) + W(\omega)\right) \cdot \left(\sum_m \lambda_m \overline{S_m(\omega)} + \overline{W(\omega)}\right) \\ &= \sum_m \sum_n \lambda_m \lambda_m S_m(\omega) \cdot \overline{S_n(\omega)} + \sum_m \lambda_m S_m(\omega) \cdot \overline{W(\omega)} \\ &\quad + \sum_m \lambda_m \overline{S_m(\omega)} \cdot W(\omega) + W(\omega) \cdot \overline{W(\omega)}. \end{aligned} \tag{16}$$

The Fourier transform $S(\omega)$ can also be written in terms of the PSD ($P(\omega)$) and phase spectral density ($\Phi(\omega)$), as

$$S(\omega) = P(\omega)^{1/2} e^{+i\Phi(\omega)}$$

which gives

$$\begin{aligned}
S_m(\omega)\cdot\overline{S_n(\omega)} &= \big(P_m(\omega)P_n(\omega)\big)^{1/2} e^{+i(\Phi_m(\omega)-\Phi_n(\omega))}\\
S_m(\omega)\cdot\overline{W(\omega)} &= \big(P_m(\omega)P_\eta(\omega)\big)^{1/2} e^{+i(\Phi_m(\omega)-\Phi_\eta(\omega))}\\
\overline{S_m(\omega)}\cdot W(\omega) &= \big(P_m(\omega)P_\eta(\omega)\big)^{1/2} e^{-i(\Phi_m(\omega)-\Phi_\eta(\omega))}\\
W(\omega)\cdot\overline{W(\omega)} &= \big(P_\eta(\omega)P_\eta(\omega)\big)^{1/2} = P_\eta(\omega)
\end{aligned}$$

where $P_\eta(\omega)$ and $\Phi_\eta(\omega)$ are PSD and phase spectral density of the noise signal. Rewriting Equation (16), we have

$$\begin{aligned}
P(\omega) = &\sum_m\sum_n \lambda_m\lambda_m \big(P_m(\omega)P_n(\omega)\big)^{1/2} e^{+i(\Phi_m(\omega)-\Phi_n(\omega))} + \sum_m \lambda_m \big(P_m(\omega)P_\eta(\omega)\big)^{1/2} e^{+i(\Phi_m(\omega)-\Phi_\eta(\omega))}\\
&+ \sum_m \lambda_m \big(P_m(\omega)P_\eta(\omega)\big)^{1/2} e^{-i(\Phi_m(\omega)-\Phi_\eta(\omega))} + P_\eta(\omega).
\end{aligned}$$

Assuming that PSD for noise is negligible compared to actual source signals, we have

$$P(\omega) = \sum_m\sum_n \lambda_m\lambda_m \big(P_m(\omega)P_n(\omega)\big)^{1/2} e^{+i(\Phi_m(\omega)-\Phi_n(\omega))}. \tag{17}$$

We know that PSD of real-valued signals is real-symmetric, hence the imaginary component in Equation (17) is zero. Hence, we have

$$P(\omega) = \sum_m\sum_n \lambda_m\lambda_m \big(P_m(\omega)P_n(\omega)\big)^{1/2} \cos\big(\Phi_m(\omega)-\Phi_n(\omega)\big).$$

$\blacksquare$

A.2 Proof of Proposition 2

Proof. The maximum likelihood estimate of $[\Omega_f, \Psi, \mathbf{X}]^T$ can be obtained by minimizing $\ell_k(\Omega_f, \Psi, X)$

$$\frac{\partial}{\partial \psi_h^{(m)}} \ell_k(\Omega_f, \Psi, \mathbf{X}) = 0$$

which leads to the following

$$P_k(h\omega_f^{(m)}) = \mathbb{P}_k(h\omega_f^{(m)}) = \left(\sum_j^M \lambda_k^{(j)} {\psi_{h_j}^{(j)}}^{1/2}\right)^2 \tag{18}$$

where

$$\psi_{h_j}^{(j)} = \begin{cases} >0 & \text{if } h_j = h\omega_f^{(m)}/\omega_f^{(j)} \in \mathbb{Z}.\\ 0 & \text{otherwise.}\end{cases}$$

If the frequency $h\omega_f^{(m)}$ is *shared* by M' sources (or the number of nonzero $\psi_{h_j}^{(j)}$ is M'), then Equation (18) becomes

$$P_k(h\omega_f^{(m)}) = \left(\sum_j^{M'} \lambda_k^{(j)} {\psi_{h_j}^{(j)}}^{1/2}\right)^2$$

If we assume the energy contribution of all the sources to be same, i.e. ${\lambda_k^{(j)}}{\psi_{h_j}^{(j)}}^{1/2} = \bar{\psi}_h$, for $j = 1, \cdots, M'$, we have

$$P_k(h\omega_f^{(m)}) = \left(M'\bar{\psi}_h\right)^2 = {M'}^2{\bar{\psi}_h}^2 = {M'}^2{\lambda_k^{(m)}}^2\psi_{h_m}^{(m)} \tag{19}$$

rearranging Equation (19), we have

$${\psi_h^{\hat{(m)}}}^{ML} = \frac{P_k(h\omega_f^{(m)})}{{M'}^2{\lambda_k^{(m)}}^2}. \tag{20}$$

Substituting the ML estimate for the energies (Equation (20)) in the negative log-likelihood (Equation (7)), we have a modified negative log-likelihood

$$\begin{aligned}\ell_k(\Omega_f, \hat{\Psi}^{ML}, \mathbf{X}) &= \ell'_k(\Omega_f, \mathbf{X}) \\ &= \sum_{\omega_j \notin \mathbb{H}(\Omega_f)} \parallel P(\omega_j) - \mathbb{P}(\omega_j) \parallel^2 \\ &\quad + \sum_{\omega_j \in \mathbb{H}(\Omega_f)} \parallel P(\omega_j) - \mathbb{P}(\omega_j) \parallel^2 \end{aligned} \tag{21}$$

where $\mathbb{H}$ is the harmonic set, which is the set of all harmonic frequencies for all sources

$$\mathbb{H}(\Omega_f) = \bigcup_m \left[\omega_f^{(m)}, 2\omega_f^{(m)}, \cdots\right]^T$$

The value of generative model $\mathbb{P}$ at the frequencies *in* the harmonic set is exactly equal to the observed PSD, hence the second term in Equation (21) goes to zero. On the other hand, the value of generative model $\mathbb{P}$ at the frequencies *not in* the harmonic set is zero, hence

$$\ell'_k(\Omega_f, \mathbf{X}) = \sum_{\omega_j \notin \mathbb{H}(\Omega_f)} \parallel P(\omega_j) - \mathbb{P}(\omega_j) \parallel^2 = \sum_{\omega_j \notin \mathbb{H}(\Omega_f)} \left(P_k(\omega_j)\right)^2. \tag{22}$$

Equation (22) is the negative log-likelihood with the constraint of Equation (20) imposed. Equation (22) implies that the modified likelihood at the ML estimate of energies is independent of the source locations

$$\ell_k(\Omega_f, \Psi^{ML}, \mathbf{X}) = \ell'_k(\Omega_f, \mathbf{X}) = \sum_{\omega_j \notin \mathbb{H}(\Omega_f)} \left(P_k(\omega_j)\right)^2 = \ell'_k(\Omega_f)$$

∎

A.3 Proof of Proposition 3

Proof. Consider a source present at an angle θ emitting a source signal $s[n]$. Using Equation (1), the received signals at the microphones are given by

$$r_p[n] = \lambda_p s[n - \tau_p] + w_p[n]$$

for $p = 1, 2$, where τ_p is the propagation delay, and λ_p is the attenuation factor. For far-field case, the distances between the source and the closely-spaced microphones will be approximately same for all microphones, hence $\lambda_1 \approx \lambda_2 = \lambda$.

Using Equation (2), the composite microphone signal for the beam angle α is given by

$$\begin{aligned} r[n] &= r_1[n] + r_2[n + t_{12}(\alpha)] \\ &= \lambda s[n - \tau_1] + \lambda s[n + t_{12}(\alpha) - \tau_2] + w_1[n] + w_2[n + t_{12}(\alpha)] \end{aligned}$$

where $t_{12}(\alpha) = t_2(\alpha) - t_1(\alpha) = d\cos(\alpha - \beta) f_s / C$ is relative sample delay. The beam energy is given by

$$\begin{aligned} B(\alpha) &= \sum_n r[n]^2 \\ &= \sum_n \left(\lambda s[n - \tau_1] + \lambda s[n + t_{12} - \tau_2] + w_1[n] + w_2[n + t_{12}]\right)^2 \\ &= \lambda^2 \sum_n s[n - \tau_1]^2 + \lambda^2 \sum_n s[n + t_{12} - \tau_2]^2 + \sum_n w_1[n]^2 + \sum_n w_2[n + t_{12}]^2 \\ &\quad + 2\lambda^2 \sum_n s[n - \tau_1] s[n + t_{12} - \tau_2] + 2 \sum_n w_1[n] w_2[n + t_{12}] \\ &\quad + 2\lambda \sum_n (w_1[n] + w_2[n + t_{12}])(s[n - \tau_1] + s[n + t_{12} - \tau_2]). \end{aligned}$$

Rewriting the above expression in terms of signal and noise autocorrelation and cross-correlation, we have

$$\begin{aligned} B(\alpha) &= \lambda^2 R_{ss}(0) + \lambda^2 R_{ss}(0) + R_{w_1 w_1}(0) + R_{w_2 w_2}(0) \\ &\quad + 2\lambda^2 R_{ss}(t_{12} - \tau_2 + \tau_1) + 2R_{w_1 w_2}(t_{12}) + 2\lambda R_{w_1 s}(-\tau_1) + 2\lambda R_{w_2 s}(t_{12} - \tau_1) \\ &\quad + 2\lambda R_{w_1 s}(t_{12} - \tau_2) + 2\lambda R_{w_2 s}(\tau_2). \end{aligned}$$

Now, assuming that the noises at the microphones are statistically same (i.e. $R_{w_1 w_1}(0) = R_{w_2 w_2}(0) = R_\eta(0)$) and the noises are uncorrelated (i.e. $R_{w_1 w_2}[m] = 0$), and the noise and signal are also uncorrelated (i.e. $R_{w_k s}[m] = 0$), we have

$$B(\alpha) = 2\lambda^2 R_{ss}(0) + 2R_\eta(0) + 2\lambda^2 R_{ss}(t_{12} - \tau_{12}).$$

Denoting $\kappa_\alpha = t_{12} - \tau_{12} = d(\cos(\alpha - \beta) - \cos(\theta - \beta)) f_s / C$ and rearranging, we have

$$B(\alpha) = 2\lambda^2 \big(R_{ss}(0) + R_{ss}(\kappa_\alpha)\big) + 2R_\eta(0).$$

■

A.4 Proof of Proposition 4

Proof. Consider a source present at an angle θ emitting a source signal $s[n]$. Using Equation (1), the received signals at the microphones are given by

$$r_p[n] = \lambda_p s[n - \tau_p] + w_p[n]$$

for $p = 1, 2, \cdots, N_{mic}$, where τ_p is the propagation delay, and λ_p is the attenuation factor. For far-field case, the distances between the source and the closely-spaced microphones will be approximately same for all microphones, hence $\lambda_1 \approx \lambda_2 \approx \cdots = \lambda$.

Using Equation (2), the composite microphone signal for the beam angle α is given by

$$\begin{aligned} r[n] &= \sum_{p=1}^{N_{mic}} r_p[n + t_{1p}(\alpha)] \\ &= \sum_{p=1}^{N_{mic}} \lambda s[n + t_{1p}(\alpha) - \tau_p] + w_p[n + t_{1p}(\alpha)] \end{aligned}$$

where $t_{1p}(\alpha) = t_p(\alpha) - t_1(\alpha) = d_{1p}\cos(\alpha - \beta_{1p}) f_s / C$ is relative sample delay between the p^{th} and 1^{st} microphone. Let's denote $\phi_p = n + t_{1p}(\alpha) - \tau_p$ and $\psi_p = n + t_{1p}(\alpha)$ for clarity and brevity. The beam energy is given by

$$\begin{aligned} B(\alpha) &= \sum_n r[n]^2 \\ &= \sum_n \left[\sum_p \lambda s[\phi_p] + w_p[\psi_p] \right]^2 \\ &= \sum_n \left[\left(\sum_p \lambda s[\phi_p] \right)^2 + \left(\sum_p w_p[\psi_p] \right)^2 + 2 \sum_p \sum_q \lambda s[\phi_p] w_q[\psi_q] \right] \\ &= \sum_n \left(\sum_p \lambda s[\phi_p] \right)^2 + \sum_n \left(\sum_p w_p[\psi_p] \right)^2 + 2 \sum_p \sum_{q, q \neq p} \underbrace{\sum_n \lambda s[\phi_p] w_q[\psi_q]}_{R_{w_q s}(\tau) = 0}. \end{aligned} \tag{23}$$

The last term is signal-noise cross-correlation which is zero for uncorrelated signal and noise. The first two term in Equation (23) are expanded to

$$\begin{aligned} \sum_n \left(\sum_p \lambda s[\phi_p] \right)^2 &= \lambda^2 \sum_n \sum_p s^2[\phi_p] + 2\lambda^2 \sum_n \sum_p \sum_{q, q \neq p} s[\phi_p] s[\phi_q] \\ &= \lambda^2 \sum_p \underbrace{\left(\sum_n s^2[\phi_p] \right)}_{R_{ss}(0)} + 2\lambda^2 \sum_p \sum_{q, q \neq p} \underbrace{\left(\sum_n s[\phi_p] s[\phi_q] \right)}_{R_{ss}(\phi_p - \phi_q)} \\ &= \lambda^2 N_{mic} R_{ss}(0) + 2\lambda^2 \sum_{p, q, p \neq q} R_{ss}(\phi_p - \phi_q) \end{aligned} \tag{24}$$

and

$$\begin{aligned} \sum_n \left(\sum_p w_p[\psi_p] \right)^2 &= \sum_n \sum_p w_p^2[\psi_p] + 2 \sum \sum_p \sum_{q, q \neq p} w_p[\psi_p] w_q[\psi_q] \\ &= \sum_p \underbrace{\left(\sum_n w_p^2[\psi_p] \right)}_{R_{w_p w_p}(0)} + 2 \sum_p \sum_{q, q \neq p} \underbrace{\left(\sum_n w_p[\psi_p] w_q[\psi_q] \right)}_{R_{w_p w_q}[\phi_p - \phi_q] = 0} \\ &= N_{mic} R_\eta(0) \end{aligned} \tag{25}$$

The second term in Equation (25) is zero due to uncorrelated noises on different microphones. Substituting Equation (24) and Equation (25) back into Equation (23), we have

$$B(\alpha) = \lambda^2 N_{mic} R_{ss}(0) + 2\lambda^2 \sum_{p,q,p\neq q} R_{ss}(\phi_q - \phi_p) + N_{mic} R_\eta(0).$$

Rearranging the terms and denoting $\kappa_{pq} = \phi_q - \phi_p = t_{pq}(\alpha) - \tau_{pq}$

$$B(\alpha) = N_{mic}(\lambda^2 R_{ss}(0) + R_\eta(0)) + 2\lambda^2 \sum_{p,q,p\neq q} R_{ss}(\kappa_{pq}). \tag{26}$$

Adding and subtracting the term $2\frac{N_{mic}(N_{mic}-1)}{2}(\lambda^2 R_{ss}(0) + R_\eta(0))$, we have

$$\begin{aligned} B(\alpha) = N_{mic}(\lambda^2 R_{ss}(0) + R_\eta(0)) + &\underbrace{2\lambda^2 \sum_{p,q,p\neq q} R_{ss}(\kappa_{pq}] + 2\frac{N_{mic}(N_{mic}-1)}{2}(\lambda^2 R_{ss}(0) + R_\eta(0))} \\ &- 2\frac{N_{mic}(N_{mic}-1)}{2}(\lambda^2 R_{ss}(0) + R_\eta(0)) \\ &= \sum_{p,q,p\neq q} \underbrace{2\left(\lambda^2 R_{ss}(\kappa_{pq}) + \lambda^2 R_{ss}(0) + R_\eta(0)\right)}_{B_{pq}(\alpha) \text{ from Proposition 3}} - N_{mic}(N_{mic}-2)(\lambda^2 R_{ss}(0) + R_\eta(0)) \\ &= \sum_{(p,q)\in\mathbb{P}} B_{pq}(\alpha) - N_{mic}(N_{mic}-2)(\lambda^2 R_{ss}(0) + R_\eta(0)) \end{aligned}$$

■

A.5 Proof of Proposition 5

Proof. Consider M sources present at angles θ_m emitting source signals $s_m[n]$, for $m = 1, 2, \cdots, M$. Using Equation (1), the received signal at the p^{th} microphone is given by

$$r_p[n] = \sum_{m=1}^{M} \lambda_{mp} s_m[n - \tau_{mp}] + w_p[n]$$

where $p = 1, 2, \cdots, N_{mic}$, τ_{mp} is the propagation delay, and λ_{mp} is the attenuation factor. For far-field case, the distances between a source and the closely-spaced microphones will be approximately same for all microphones, hence $\lambda_{m1} \approx \lambda_{m2} \approx \cdots = \lambda_m$.

Using Equation (2), the composite microphone signal for the beam angle α is given by

$$\begin{aligned} r[n] &= \sum_{p=1}^{N_{mic}} r_p[n + t_{1p}(\alpha)] \\ &= \sum_{p=1}^{N_{mic}} \sum_{m=1}^{M} \lambda_m s_m[n + t_{1p}(\alpha) - \tau_{mp}] + w_p[n + t_{1p}(\alpha)] \end{aligned}$$

where $t_{1p}(\alpha) = t_p(\alpha) - t_1(\alpha) = d_{1p}\cos(\alpha - \beta_{1p})f_s/C$ is relative sample delay between the p^{th} and 1^{st} microphone. Let's denote $\phi_{mp} = n + t_{1p}(\alpha) - \tau_{mp}$ and $\psi_p = n + t_{1p}(\alpha)$ for clarity and brevity. The beam energy is given by

$$\begin{aligned}
B(\alpha) &= \sum_n r[n]^2 \\
&= \sum_n \left[\sum_p \sum_m \lambda_m s_m[\phi_{mp}] + w_p[\psi_p]\right]^2 \\
&= \sum_n \left[\left(\sum_p \sum_m \lambda_m s_m[\phi_{mp}]\right)^2 + \left(\sum_p w_p[\psi_p]\right)^2 + 2\sum_p \sum_q \sum_m \lambda_m s_m[\phi_{mp}] w_q[\psi_q]\right] \\
&= \sum_n \left(\sum_p \sum_m \lambda_m s_m[\phi_{mp}]\right)^2 + \underbrace{\sum_n \left(\sum_p w_p[\psi_p]\right)^2}_{N_{mic}R_\eta(0) \text{ using Equation (25).}} \\
&\quad + 2\sum_p \sum_q \sum_m \underbrace{\sum_n \lambda_m s_m[\phi_{mp}] w_q[\psi_q]}_{R_{w_q s_m}(\tau)=0}.
\end{aligned} \tag{27}$$

The first term in Equation (27) is expanded to

$$\begin{aligned}
\sum_n \left(\sum_p \sum_m \lambda_m s_m[\phi_{mp}]\right)^2 &= \sum_n \left(\sum_m \sum_p \lambda_m s_m[\phi_{mp}]\right)^2 \\
&= \sum_n \left(\sum_m \left(\sum_p \lambda_m s_m[\phi_{mp}]\right)^2 + 2\sum_{m_1}\sum_{m_2} \lambda_{m_1} s_{m_1}[\phi_{m_1 p}]\lambda_{m_2} s_{m_2}[\phi_{m_2 p}]\right) \\
&= \sum_m \left(\sum_n \left(\sum_p \lambda_m s_m[\phi_{mp}]\right)^2\right) + 2\sum_{m_1}\sum_{m_2} \underbrace{\sum_n \lambda_{m_1} s_{m_1}[\phi_{m_1 p}]\lambda_{m_2} s_{m_2}[\phi_{m_2 p}]}_{R_{s_{m_1}s_{m_2}}(\tau)=0} \\
&= \sum_m \underbrace{\left(\sum_n \left(\sum_p \lambda_m s_m[\phi_{mp}]\right)^2\right)}_{\text{substitute from Equation (24)}} \\
&= \sum_m \lambda_m^2 \left(N_{mic}R_{s_m s_m}(0) + \sum_{p,q,p\neq q} R_{s_m s_m}(\phi_{mp} - \phi_{mq})\right)
\end{aligned} \tag{28}$$

Denoting $\kappa_{mpq} = \phi_{mq} - \phi_{mp} = t_{pq}(\alpha) - \tau_{mpq}$, and substituting Equation (28) in Equation (27), we have

$$B(\alpha) = \sum_m \lambda_m^2 \left(N_{mic}R_m(0) + \sum_{p,q,p\neq q} R_m(\kappa_{mpq})\right) + N_{mic}R_\eta(0)$$

Adding and subtracting the term $\sum_m N_{mic} R_\eta(0)$, we have

$$\begin{aligned}
B(\alpha) &= \sum_m \lambda_m^2 \left(N_{mic} R_m(0) + \sum_{p,q,p \neq q} R_m(\kappa_{mpq}) \right) + \sum_m N_{mic} R_\eta(0) - \sum_m N_{mic} R_\eta(0) + N_{mic} R_\eta(0) \\
&= \sum_m \underbrace{\lambda_m^2 \left(N_{mic} R_m(0) + \sum_{p,q,p \neq q} R_m[\kappa_{pq}] \right) + N_{mic} R_\eta(0)}_{B_m(\alpha),\text{ using Equation (26)}} - M N_{mic} R_\eta(0) + N_{mic} R_\eta(0) \\
&= \sum_m B_m(\alpha) - N_{mic}(M-1) R_\eta(0)
\end{aligned}$$

∎