ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413765

MISALIGNMENT RECOGNITION IN ACOUSTIC SENSOR NETWORKS USING A SEMI-SUPERVISED SOURCE ESTIMATION METHOD AND MARKOV RANDOM FIELDS

Gabriel F Miller, Andreas Brendel, Walter Kellermann

Friedrich-Alexander-Universität, Erlangen-Nürnberg
Multimedia Communications and Signal Processing Lab
Cauerstr. 7, D-91058 Erlangen, Germany

Sharon Gannot

Bar-Ilan University
Faculty of Engineering
Ramat-Gan, Israel

ABSTRACT

In this paper, we consider the problem of acoustic source localization by acoustic sensor networks (ASNs) using a promising, learning-based technique that adapts to the acoustic environment. In particular, we look at the scenario when a node in the ASN is displaced from its position during training. As the mismatch between the ASN used for learning the localization model and the one after a node displacement leads to erroneous position estimates, a displacement has to be detected and the displaced nodes need to be identified. We propose a method that considers the disparity in position estimates made by leave-one-node-out (LONO) sub-networks and uses a Markov random field (MRF) framework to infer the probability of each LONO position estimate being aligned, misaligned or unreliable while accounting for the noise inherent to the estimator. This probabilistic approach is advantageous over naïve detection methods, as it outputs a normalized value that encapsulates conditional information provided by each LONO sub-network on whether the reading is in misalignment with the overall network. Experimental results confirm that the performance of the proposed method is consistent in identifying compromised nodes in various acoustic conditions.

Index Terms— Acoustic manifold learning, failure detection, Gaussian process, Markov random fields, sound source localization.

1. INTRODUCTION

Sound source localization is a topic that has been covered in great detail and remains a burgeoning field of study [1–10], see [11] for an overview of the state of the art. Especially, smart-home technology drove the need for robust and efficient localization methods in acoustic sensor networks (ASNs) [12,13]. While in the past, traditional localization methods typically relied on physics-based models [2–4], there has been a growing interest in localizing acoustic sources using learning-based methods whereby position estimates are obtained directly from previously learned knowledge about a given acoustic environment. These methods have been shown to be effective, particularly in adverse acoustic conditions [1, 5–7, 10, 14] as long as the parameters used for training remain static. For example, when localizing sources in a smart-home environment, many of the underlying characteristics of the room remain essentially unchanged (e.g., the room dimensions and reverberation time). This means the variability regarding the acoustic transfer functions, which are typically represented in a high-dimensional feature space, can be mostly attributed to the source position. This lends credibility to the use of learning-based methods where these static qualities can be captured during a training phase.

Due to the difficulty in acquiring labelled data, a semi-supervised method based on a small labelled and a large unlabelled data set is generally employed. Unlabelled data, which are easy to obtain, are used together with a few labelled ‘anchors’ to train models for acoustic source localization [9]. In [15], a semi-supervised approach was employed for source localization using a relative transfer function (RTF)-based feature vector, which measures the relation between the acoustic paths from a sound source to two different microphones. Thus, by leveraging unlabelled data, a more robust localizer is achieved. In this study the scenario considered was limited to a single microphone system in a static environment with white Gaussian noise input. Subsequently, in [10], the semi-supervised inference approach, based on Gaussian process (SSGP) on multiple manifolds, was further developed and adapted to localize a speech source using a multi-microphone system, again based on a dense grid of RTFs [10, 16, 17].

However, if the array constellation, e.g., the position of one or more nodes, changes relative to the training stage, the usefulness of the learned model becomes uncertain. Such a displacement of a node may be caused by the user changing the position of, e.g., a smart speaker in a smart home environment. In our work, we adopt the SSGP method and consider the scenario where any given microphone node can be moved and the sources are assumed to be static for at least a short period of time. The detrimental effect of an array movement on the localization error can be observed in Fig. 1, where the error almost doubles with only a small shift of a random node in the network. We are thus posed with the problem of determining if a node is moving, and specifically determining which of the nodes is moving. In order to address both issues, we consider a technique recently introduced in the field of robotics for recognizing sensor misalignment [18]. The authors in [18] utilize Markov random fields (MRFs) with fully connected latent variables (FCLVs) to measure the probability of misalignment of a sensor network based on individual sensor readings and a ground truth mapping of a given room [19, 20]. Recognition of misalignment is needed in [18] to determine whether differences in measurements sampled over time should be attributed to actual changes or due to inherent noise.

Rather than taking each sensor signal independently, we look at the so-called leave-one-node-out (LONO) sub-network position estimates (with each sub-network containing all but one node) obtained via the SSGP method. We then use the differences between the position estimates of the LONO sub-networks as input to the MRF model (note, for our considerations in this paper the sound source is as-

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 282835863 – within the Research Unit FOR2457 “Acoustic Sensor Networks” and from the European Union’s Horizon 2020 Research and Innovation Programme, Grant Agreement No. -871245.

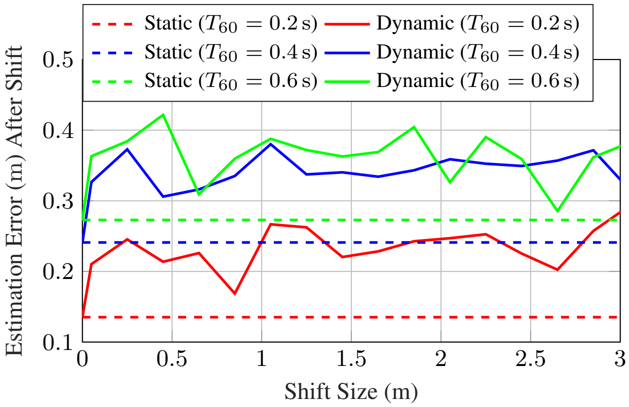

Fig. 1. Mean error of the SSGP localizer for the scenario of Fig. 2 with different reverberation times T_{60}, comparing scenarios without node movement (dotted lines) to scenarios when one node in the network moved (random node shifted from its learned position) as can be seen in solid lines.

sumed to be short-time static). Eventually our model outputs posterior probabilities per LONO sub-network for belonging to one of the following latent states: aligned, misaligned or unreliable. These posteriors are used to indicate both the probability of movement in the network, and also allows for inference of which node moved.

2. REVIEW OF THE SSGP SOURCE LOCALIZATION TECHNIQUE

We now briefly review the SSGP source localization method (see [10] for details) and consider a speech signal in the Short-Time Fourier Transform (STFT) domain, $S(\tau, k)$, at frame index τ, frequency index k, received at a given node m, and emitted from position $\mathbf{q}$. We then model the signal received at node m as

$$Y_i^m(\tau, k) = A_i^m(\tau, k, \mathbf{q})\, S(\tau, k) + U_i^m(\tau, k)\,, \tag{1}$$

with i specifying the ith microphone in the mth node. Additionally, $A_i^m(\tau, k, \mathbf{q})$ is the acoustic transfer function (ATF) relating the sound source originating at position $\mathbf{q}$ to the ith microphone, and $U_i^m(\tau, k)$ is the STFT-domain representation of an additive noise signal which corrupts the measurement. The spatial information required for localizing a source at position $\mathbf{q}$ is embedded in the ATF, and is independent of the source signal. Rather than extracting the ATF we use the aforementioned RTF feature vector, $\mathbf{h}^m$ (defined as the ratio of two ATFs [21]) as it is easier to acquire in practice and is equally informative for the proposed localization method.

In order to determine the position of an unknown source, we first define $\mathbf{q}_t = [q_{t,x}, q_{t,y}, q_{t,z}]^\top$ as the unknown 'test' position to be inferred from an estimated RTF sample (e.g., by least squares), $\mathbf{h}_t^m$, which relates the unknown source position $\mathbf{q}_t$ to node m, assuming that each of the M network nodes has only two microphones. For training the SSGP estimator, a set of n_D sound sources is used where n_D is the number of training points, from which n_U are measured RTFs and n_L are measured RTFs with associated source positions serving as labels ($n_L + n_U = n_D$). Each Cartesian coordinate $p_{d,a}$, $a \in \{x, y, z\}$ of a training position, $\mathbf{p}_d \in \mathbb{R}^3$ is related to an RTF sample $\mathbf{h}_d^m$ measured at node m via an unknown functional relationship. Moreover, we assume that the coordinates of all n_D labelled and unlabelled training positions captured by vectors $\mathbf{p}_{D,a} = \left[p_{1,a} \ldots, p_{n_D,a}\right]^\top$, are all jointly Gaussian, and their relation to the source positions can be determined by the posterior mean function of corresponding Gaussian distributions. We will now discuss how we utilize the labelled and unlabelled RTF training samples $\{\mathbf{h}_d^m\}_{d=1}^{n_D}$ $\left(\forall m \in \{1, \ldots, M\}\right)$ in order to identify an unknown source position from its corresponding measured RTF $\mathbf{h}_t^m$. We will omit the dependency on the coordinate $a \in \{x, y, z\}$ in the following for conciseness.

In order to localize a sound source, RTFs obtained at each node are related to those obtained at every other node in the ASN, which is summarized via the kernel-based covariance matrix, $\boldsymbol{\Sigma}_L$, with each element representing a pairwise affinity between two RTF samples. In particular, we express an element of the covariance matrix which relates two labelled source positions, indexed by i and j as

$$(\boldsymbol{\Sigma}_L)_{i,j} = \frac{1}{M^2}\sum_{d=1}^{n_D}\sum_{q=1}^{M}\sum_{w=1}^{M} k_q\left(\mathbf{h}_i^q, \mathbf{h}_d^q\right) k_w\left(\mathbf{h}_j^w, \mathbf{h}_d^w\right). \tag{2}$$

Here, $\mathbf{h}_i^m$, $\mathbf{h}_j^m$ are RTF samples from the set of labelled RTFs $\mathcal{H}_L = \{\mathbf{h}_i^1, \ldots, \mathbf{h}_i^M\}_{i=1}^{n_L}$, and $k_m(\mathbf{h}_i^m, \mathbf{h}_j^m)$ is a conventional pairwise Gaussian kernel function, $k_m : \mathcal{M}_m \times \mathcal{M}_m \rightarrow \mathbb{R}_+$ with:

$$k_m\left(\mathbf{h}_i^m, \mathbf{h}_j^m\right) = \exp\left\{-\frac{\left\|\mathbf{h}_i^m - \mathbf{h}_j^m\right\|_2^2}{\varepsilon_m}\right\}, \tag{3}$$

where $\mathcal{M}_m$ denotes a manifold corresponding to node m, and ε_m is a parameter defining the width of the kernel [22]. Similarly, we can define an element in the test covariance vector, $\boldsymbol{\Sigma}_{Lt} \in \mathbb{R}^{n_L}$, used for inferring the position of an unknown source element

$$(\boldsymbol{\Sigma}_{Lt})_i = \frac{1}{M^2}\sum_{d=1}^{n_D}\sum_{q=1}^{M}\sum_{w=1}^{M} k_q\left(\mathbf{h}_i^q, \mathbf{h}_d^q\right) k_w\left(\mathbf{h}_t^w, \mathbf{h}_d^w\right). \tag{4}$$

The unknown position, $\mathbf{q}_t$, can thus be estimated coordinate-wise via the conditional mean $\mathcal{E}\left\{q_t | \mathbf{p}_L, \mathcal{H}_L, \{\mathbf{h}_t^m\}_{m=1}^M\right\}$, where $\mathbf{p}_L \in \mathbb{R}^{n_L}$ is the vector containing one of the coordinates of the labelled training positions. The distribution of all source positions (known and unknown) is defined over the concatenation of all coordinates of the labelled training positions $\mathbf{p}_L$ and the coordinate to be estimated q_t

$$\begin{bmatrix} \mathbf{p}_L \\ q_t \end{bmatrix} \Bigg| \mathcal{H}_L \backsim \mathcal{N}\left(\mathbf{0}_{n_L+1}, \begin{bmatrix} \boldsymbol{\Sigma}_L + \sigma^2 \mathbf{I}_{n_L} & \boldsymbol{\Sigma}_{Lt} \\ \boldsymbol{\Sigma}_{Lt}^\top & \Sigma_t \end{bmatrix}\right), \tag{5}$$

where Σ_t is the variance of q_t, σ^2 is the variance associated with the accuracy of the labels, $\mathbf{I}_{n_L}$ is the $n_L \times n_L$ identity matrix and $\mathbf{0}_{n_L+1}$ is an all-zero vector of length $n_L + 1$. Thus, we estimate the position of an unknown source by the conditional mean associated with (5):

$$\hat{q}_t = \mathcal{E}\left\{q_t | \mathbf{p}_L, \mathcal{H}_L, \{\mathbf{h}_t^m\}_{m=1}^M\right\} = \boldsymbol{\Sigma}_{Lt}^\top \left(\boldsymbol{\Sigma}_L + \sigma^2 \mathbf{I}_{n_L}\right)^{-1} \mathbf{p}_L. \tag{6}$$

An example of the localization scenario is shown in Fig. 2 (detailed room specifications can be found in Sec. 4.).

3. MISALIGNMENT DETECTION

In our scenario, MRFs are used to determine if a node in an ASN is displaced, and also determine which node moved. MRFs provide a convenient and consistent way of modeling context dependent entities and can be implemented in a local and massively parallel manner [19, 23]. MRFs are especially useful for inference if

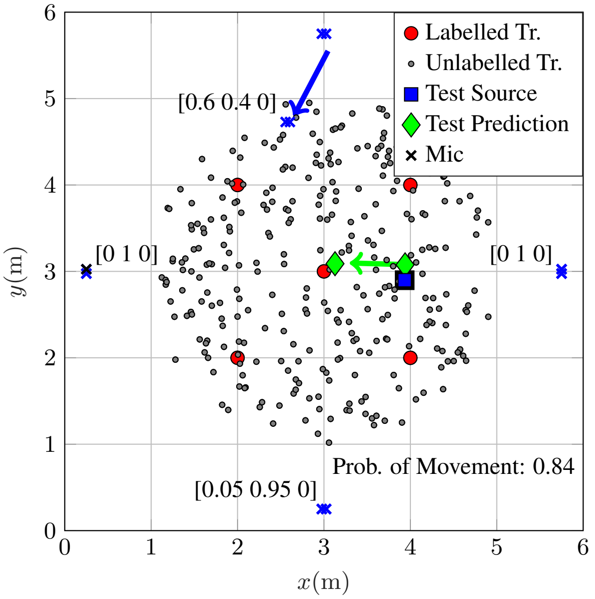

Fig. 2. Misaligned scenario with example of proposed detection method. Green arrow indicates how the prediction of an acoustic source changes based on the translation of a random node (blue arrow). Values in brackets next to nodes indicate probability of a LONO sub-network being aligned, misaligned, or unreliable where the referenced node is the one left out. Higher probabilities of alignment indicate the node left out is likely compromised.

a priori probability functions are given for the latent variables governing the observations. The observed quantities we use as input to the MRF model are the SSGP localization estimates of M LONO sub-networks. These estimates are compared to the localization estimates recorded before movement for the mth LONO sub-network via the Euclidean distance

$$e_m = \|\mathbf{q}_m - \hat{\mathbf{q}}_m\|_2. \tag{7}$$

Note that, here, $\mathbf{q}_m$ refers to position estimates recorded by a given LONO sub-network before movement occurs, and $\hat{\mathbf{q}}_m$ refers to the estimate after movement. While in dynamic scenarios with moving sources these distinctions will not be so clear, in the (short-time) static scenario assumed here, they are useful for the desired analysis. The latent variables are given by the errors made by a LONO in a given acoustic environment, which is dependent on the room itself and, consequently, on the variability of the SSGP localizer. Therefore, in practice, the considered MRF model (as detailed in Sections 3.1, 3.2) compares the difference obtained from each LONO sub-network using a message passing scheme and incorporates prior information regarding the expected deviation between the LONO estimates by a general localization error distribution. This distribution is acquired by simulating localization estimates of LONO sub-networks after a random array in the ASN is displaced in a random direction (including displacements by 0, 0.65, 0.85 and 1.5 m) and comparing it with the ground truth position of the source (ground truth choices include 0.1, 0.75 and 1.45 m). The output of the model is a probability indicating if the network is in alignment. We assume the latent variables to be FCLVs to ensure the difference in estimation measured by each LONO sub-network is compared with the difference measured by every other LONO sub-network. Additionally, we assume that only one node in the network is moving at a time, therefore, the sub-network with the smallest probability of movement as determined by the MRF model would be the one that did not contain the moved node, thus allowing us to infer which particular node was moving. For this inference, we consider the posterior probability output by the MRF that a given LONO sub-network m, is of one of the following latent classes: aligned, misaligned or unreliable.

3.1. Likelihood distributions of estimation errors

For describing the estimation method, we first introduce the latent posterior probabilities, $\mathbb{P}\left(\mathbf{z}_m \mid \mathbf{e}\right)$, where $\mathbf{z}_m = [z_{m,1}, z_{m,2}, z_{m,3}]^\top$ is an indicator vector of binary variables with each variable indicating whether a given LONO sub-network belongs to a given latent class, and where $\mathbf{e} = [e_1, \ldots, e_M]^\top$ is the vector containing the errors e_m (see (7)) from all LONO sub-networks.

We also define the prior error distributions of each latent class, which were found empirically from observed errors (by simulating a localization scenario for various noise levels and T_{60} values). For the aligned case, $\mathbf{z}_m = [1, 0, 0]$, we found the error distribution to be similar to a half-normal distribution with variance σ^2_{align} [18,24],

$$\mathbb{P}\left(e_m \mid \mathbf{z}_m = [1,0,0], \sigma^2_{\text{align}}\right) = 2\mathcal{N}\left(e_m; 0, \sigma^2_{\text{align}}\right), e_m \geq 0. \tag{8}$$

For the misaligned case, an exponential distribution with parameter λ was chosen, $\mathbf{z}_m = [0, 1, 0]$,

$$\mathbb{P}\left(e_m \mid \mathbf{z}_m = [0,1,0], \lambda\right) = \frac{\lambda \exp\{-\lambda\, e_m\}}{1 - \exp\{-\lambda\, e_{\max}\}}, \tag{9}$$

and a uniform distribution for unreliable observations, $\mathbf{z}_m = [0, 0, 1]$,

$$\mathbb{P}\left(e_m \mid \mathbf{z}_m = [0,0,1]\right) = \text{unif}\left(0, e_{\max}\right), \tag{10}$$

where, $e_{\max}$ references the maximum localization error. A uniform distribution is assigned to the unreliable class (analogous to the assumption made in [18]) to reflect the uninformative character of this class, as we assume that the movement of the nodes in the network cannot be predicted. The hyper-parameters corresponding to the aligned (normal) and misaligned (exponential) were estimated via a random grid search [25].

3.2. Latent class estimation and failure detection

As noted, we make the FCLV assumption which allows us to consider the viewpoint of every LONO sub-network in calculating the set of latent posterior probabilities for a specific LONO sub-network. In particular, every set of latent variables associated with a LONO sub-network receives messages from all other nodes and their corresponding set of variables to initialize the marginal posterior probabilities, which is calculated as

$$\mathbb{P}\left(\mathbf{z}_m \mid \mathbf{e}\right) = \frac{1}{Z}\mathbf{l}_m \odot \prod_{\substack{m'=1 \\ m' \neq m}}^{M} \boldsymbol{\mu}_{m' \to m}\left(\mathbf{z}_m\right). \tag{11}$$

Here, Z is a normalizing factor, $\odot$ is the Hadamard product and $\mathbf{l}_m$ is a likelihood vector

$$\mathbf{l}_m = \left[\mathbb{P}\left(e_m \mid z_{m,1}\right), \mathbb{P}\left(e_m \mid z_{m,2}\right), \mathbb{P}\left(e_m \mid z_{m,3}\right)\right]^\top. \tag{12}$$

The message from m'th to the mth LONO sub-network is denoted as

$$\boldsymbol{\mu}_{m'\to m}\left(\mathbf{z}_m\right) = \psi_{m',m}\left(\mathbf{z}_{m'},\mathbf{z}_m\right)\mathbf{l}_{m'}. \tag{13}$$

In this case, $\psi_{m',m}\left(\mathbf{z}_{m'},\mathbf{z}_m\right)$ is the transition probability from state m' to m and is an element of the transition matrix $\boldsymbol{\psi} \in \mathbb{R}_{+}^{3\times 3}$. The matrix is optimized using an iterative proportional fitting procedure, again based on empirical localization errors [19, 26].

Finally, after each node receives initial messages from all other nodes, messages are continually passed around until convergence to the maximum likelihood posterior.

With the posteriors for each LONO sub-network, we obtain the probability of misalignment in the overall network based on the average posterior probabilities of misalignment for all sub-networks:

$$p_{\text{failure}} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{P}\left(z_{m,2}\mid\mathbf{e}\right). \tag{14}$$

Then, the criterion $p_{\text{failure}} \geq p_{\text{thresh}}$ with the user-defined threshold p_{thresh} is used for detecting node movement. A misaligned scenario is illustrated in Fig. 2 where a node is displaced by one meter. Values in brackets next to each node indicate the probability of a LONO sub-network being aligned, misaligned, or unreliable where the referenced node is the one left out. Higher probabilities of alignment indicate that the node left out is more probable to have moved.

4. EVALUATION

We present a simulation study showing the efficacy of the proposed method. After describing the experimental setup we discuss the results obtained from Monte-Carlo simulations.

We consider a room of size 6 m × 6 m × 3 m with four nodes uniformly spaced in a square (see Fig. 2), each comprising two microphones spaced 5 cm apart. The Region of Interest (RoI) was chosen to be in the center of the node network and within a 2 m radius from the center of the room. In total, we simulated five labelled sources and 300 unlabelled sources to generate RTFs, whereby each unlabelled point was randomly chosen from a uniform 2D distribution within the RoI. White noise convolved with simulated room impulse responses (RIRs) [27] has been used for training.

The SSGP parameters were optimized via an ML estimation (see [10] for details) for varying noise levels and T_{60}. This was done by drawing at random speech signals from a database of English speakers [28], again convolving them with simulated RIRs, randomizing the position of the source and comparing the positional estimates to the ground truth position. The parameters of the MRF model, σ^2_{align}, λ, and $e_{\max}$ were chosen via a random grid search whereby a room environment was simulated and arrays were randomly shifted. The optimal parameters were then chosen based on the detector's ability to recognize movement for a range of probability thresholds. Care was taken in choosing these thresholds, as extremely small thresholds result in a high number of false positives as even a movement occurring with only a small probability will be declared movement by the MRF model, and without loss of generality, large thresholds result in a large number of false negatives. Thus the threshold was incremented (from 0 to 1 by increments of 0.05 m) to balance the range of possible outcomes.

For the results in Fig. 3, the movement detection scenario was simulated over 100 trials per shift of a randomly chosen node, shifted in a random direction, and for a range of reverberation levels. The movement detection probability increases with the size of the displacement, and is largely independent of the T_{60} level. We attribute the fact that the curves are not monotonic to the variance inherent to the SSGP technique.

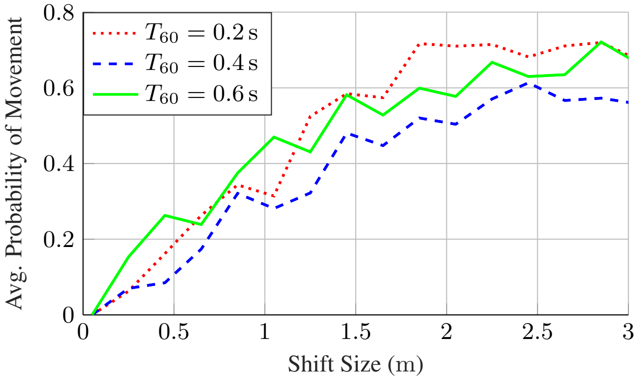

Fig. 3. Output from the MRF-based detector for incremental shifts of a random node and varying T_{60} with 100 trials per shift and T_{60}.

In order to test the robustness of the proposed algorithm, we compare it to a naïve detector that uses the LONO positional estimates directly as a way of indicating movement. Thus the naïve detector will indicate movement occurred if the deviation for a given LONO sub-network is greater than the average of the other three LONO estimates and this difference exceeds some threshold. Thereby the thresholds were chosen based on the average difference between a LONO sub-network excluding the shifted node and the mean estimates of the other three. We found (as indicated in Table 1) that the MRF-based detector performs better for most T_{60} levels with respect to the area under the receiver operating curve (AUC) [29]. Note that the basis of the decision of the naïve detector is essentially the input to the MRF-based detector(i.e., $\mathbf{e}$). Thus, for the most part we observe improvement achieved by incorporating the prior information regarding the error distributions rather than the mean of the errors.

T_{60} [s]	0.2	0.4	0.6
Naïve	0.71	0.62	0.78
MRF	0.84	0.82	0.78

Table 1. AUCs reported for the LONO sub-network estimation comparison and the MRF-based detector at varying T_{60}.

5. CONCLUSION

In this paper, we proposed a method for consistently identifying situations where moving sensor network nodes render source localization estimates questionable or useless. Specifically we considered the problem of detecting the movement of a microphone node in a network. The proposed probabilistic MRF-based algorithm determines whether a network of nodes is aligned with the previously learned configuration by leveraging prior information on the error distribution of an SSGP localization technique. The benefit of the MRF model was demonstrated by comparison to an estimate that relied directly on the relative difference in positional estimates by different sub-networks of nodes. In particular, we showed that the MRF-based detector outputs a movement indicator that scales commensurate with the size of disruption in the network, and one that is consistent across varying T_{60}. As of now the algorithm assumes a static source, and its application to a moving sound source is planned as future work.

6. REFERENCES

[1] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Speaker Tracking on Multiple-Manifolds with Distributed Microphones," *Latent Variable Analysis and Signal Separation*, vol. 10169, pp. 59–67, Feb, 2017.

[2] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, p. 276–280, 1986.

[3] R. Roy and T. Kailath, "ESPRIT-Estimation of Signal Parameters via Rotational Invariance Techniques," *IEEE International Conference Acoustic, Speech, and Signal Processing*, vol. 37, no. 7, p. 984–995, Jul, 1989.

[4] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Antennas and Propagation*, vol. ASSP-24, no. 4, p. 320–327, Aug, 1976.

[5] A. Deleforge and R. Horaud, "2D Sound-Source Localization on the Binaural Manifold," *Proc. IEEE Inernational Workshop Machine Learning and Signal Processing*, pp. 1–6, Sep, 2012.

[6] A. Deleforge, F. Forbes, and R. Horaud, "Variational EM for Binaural Sound-Source Separation and Localization," *Proc. IEEE Int. Workshop Machine Learning Signal Processing*, pp. 76–80, 2013.

[7] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds," *International Journal of Neural Systems*, vol. 25, no. 1, 2015.

[8] T. May, S. van de Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, Jan, 2011.

[9] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-Supervised Sound Source Based on Manifold Regularization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 8, pp. 1393–1407, Aug, 2016.

[10] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised source localization on multiple-manifolds with distributed microphones," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 7, pp. 1477–1491, Jul, 2017.

[11] C. Evers, H. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. Naylor, and W. Kellermann, "The LOCATA Challenge: Acoustic Source Localization and Tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[12] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A Survey of Sound Source Localization Methods in Wireless Acoustic Sensor Networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–24, 2017.

[13] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing Multiple Audio Sources in a Wireless Acoustic Sensor Network," *Signal Processing*, vol. 107, pp. 54–67, Feb. 2015.

[14] A. Brendel and W. Kellermann, "Distributed Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 61–75, 2019.

[15] B. Laufer, R. Talmon, and S. Gannot, "Relative Transfer Function Modeling for Supervised Source Localization," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

[16] R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth, "Manifold Gaussian Processes for Regression," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 3338–3345, July 2016.

[17] M. Bianco, G. P., and S. Gannot, "Semi-Supervised Source Localization with Deep Generative Modeling," in *30th Machine Learning for Signal Processing (MLSP)*, (Aalto University, Espoo, Finland), Sept. 2020.

[18] N. Akai, L. Morales Yoichi, T. Hirayama, and H. Murase, "Misalignment Recognition Using Markov Random Fields with Fully Connected Latent Variables for Detecting Localization Failures," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3955–3962, Jul, 2019.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006.

[20] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," *CoRR*, vol. abs/1210.5644, 2012.

[21] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[22] M. Genton, "Classes of Kernels for Machine Learning: A Statistics Perspective," *Journal of Machine Learning Research*, vol. 2, pp. 299–312, 01 2001.

[23] S. Li, "Markov Random Field Modeling in Computer Vision," *Springer, Tokyo*, 1995.

[24] R. H. Byers, "Half-Normal Distribution," *Encyclopedia of Biostatistics*, 2005.

[25] K. Ensor and P. Glynn, "Stochastic Optimization via Grid Search," *Mathematics of Stochastic Manufacturing Systems: AMS-SIAM Summer Seminar in Applied Mathematics*, p. 89 – 100, 1997.

[26] S. E. Fienberg and M. M. Meyer, "Iterative Proportional Fitting," Tech. Rep. 270, Department of Statistics, Carnegie-Mellon University, June 1981.

[27] E. Habets, *Room Impulse Response Generator*. International Audio Laboratories, Am Wolfsmantel 33, 91058 Erlangen, Germany, Sep, 2010. https://github.com/ehabets/RIR-Generator.

[28] D. Povey, "ST-AEDS-20180100_1, Free ST American English corpus," 2018.

[29] A. P. Bradley, "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognition*, vol. 30, p. 1145–1159, July 1997.