MULTI-VIEW SOURCE LOCALIZATION BASED ON POWER RATIOS

Bracha Laufer-Goldshtein[1]*, Ronen Talmon*[2]*, Israel Cohen*[2] *and Sharon Gannot*[1]

[1] Faculty of Engineering
Bar-Ilan University
Ramat-Gan, 5290002, Israel
{bracha.laufer,sharon.gannot}@biu.ac.il

[2] Viterbi Faculty of Electrical Engineering
The Technion - Israel Institute of Technology
Haifa, 3200003, Israel
{ronen,icohen}@ee.technion.ac.il

ABSTRACT

Despite attracting significant research efforts, the problem of source localization in noisy and reverberant environments remains challenging. Novel learning-based methods attempt to solve the problem by modelling the acoustic environment from the observed data. Typically, appropriate feature vectors are defined, and then used for constructing a model, which maps the extracted features to the corresponding source positions. In this paper, we focus on localizing a source using a distributed network with several arrays of unidirectional microphones. We introduce new feature vectors, which utilize the special characteristic of unidirectional microphones, receiving different parts of the reverberated speech. The new features are computed locally for each array, using the power-ratios between its measured signals, and are used to construct a local model, representing the unique view point of each array. The models of the different arrays, conveying distinct and complementing structures, are merged by a Multi-View Gaussian Process (MVGP), mapping the new features to their corresponding source positions. Based on this unifying model, a Bayesian estimator is derived, exploiting the relations conveyed by the covariance terms of the MVGP. The resulting localizer is shown to be robust to noise and reverberation, utilizing a computationally efficient feature extraction.

Index Terms— source localization, unidirectional microphones, supervised learning, Gaussian process

1. INTRODUCTION

Source localization is a core problem in speech processing, with various applications, such as teleconferencing systems [1], robot audition [2], speech enhancement and separation [3], just to name a few. Numerous methods were proposed over the years based on the spatial information inferred from the measurements of several microphones [4]. A substantial number of methods are based on time difference of arrival (TDOA) estimates between pairs of microphones, which are combined to perform the actual localization [5–7]. Some localization schemes utilize beamforming techniques [8] or subspace analysis [9–13]. Other existing approches incorporate spatial sparsity considerations [14, 15].

Most of the above methods demonstrate limited performance in adverse conditions, i.e., in the presence of noise and reverberations. These complex scenarios are difficult to analyse, and accurate definitive geometrical or physical models do not always exist. Novel learning based methods tackle the problem by either supervised or unsupervised techniques, aiming to learn the spatial characteristics of the acoustic environment directly from the data [16–19].

Recently, we have presented several localization and tracking algorithms based on semi-supervised learning [20–23]. In these algorithms the first step is to extract features from the measured signals in several microphones, representing the *acoustic fingerprint* of each source position. The extracted features are then mapped to the corresponding source positions using manifold-regularized optimization techniques [21], as well as using Bayesian regression with Gaussian processes [22,23]. Typically the high-dimensional features used are based on relative transfer function (RTF) estimates, representing the corresponding acoustic channels. Localization results, compared to well-known TDOA-based localization schemes [5, 24], show superiority of the proposed methods in moderate and high reverberation conditions, whereas in low reverberations the TDOA-based approaches are slightly preferable [21, 22]. Moreover, robustness to background noise was not explicitly addressed.

In this paper, we deal with distributed arrays of unidirectional microphones. We propose new features, specially tailored to work with unidirectional microphones, that can serve as a simple, yet powerful alternative to the intricate high-dimensional RTFs. The new features are based on ratios between temporal energies measured by the different microphones in each array [25]. We show that the ratio-based features are advantageous over the RTF-based features for their low-computational complexity as well as for their better localization accuracy in low reverberation conditions or in the presence of high noise levels.

The outline of the proposed algorithm is as follows. In the first step, the features are computed for several spatially distributed microphone arrays. Each array is characterized by a unique acoustic fingerprint, representing its local *view point*. The different fingerprints are then incorporated into a definition of a Multi-View Gaussian Process (MVGP), mapping the features of the different arrays to their corresponding source positions. In the last step, a Bayesian localizer is derived, yielding an estimation based on the relations defined by the MVGP.

2. PROBLEM FORMULATION

We consider a single source located at position $\mathbf{p} = [p_x, p_y, p_z]^T$ in a reverberant enclosure. The source signal $s(t)$ is contaminated by noise, and is measured by M distributed nodes, each of which consists of an array with J microphones. The measured signal in the jth microphone of the mth array is given by:

$$y_j^m(t, \mathbf{p}) = h_j^m(t, \mathbf{p}) * s(t) + v_j^m(t) \tag{1}$$

[1] This work is supported by the Adams Foundation of the Israel Academy of Sciences and Humanities .

where $h_j^m(t,\mathbf{p})$ is the acoustic impulse response (AIR) relating the source at position $\mathbf{p}$ and the jth microphone in the mth node, and $v_j^m(t)$ is the corresponding noise signal. In this paper, we assume that each microphone array comprises unidirectional elements, each directed at a different angle. In the sequel, we utilize the spatial characteristic of this type of array to extract simple features, which vary smoothly with respect to the source position.

Suppose we are given a training set of N prerecorded measurements in various known source positions $\{\bar{\mathbf{p}}_n\}_{n=1}^N$ in the enclosure of interest. The measured training positions deviate from the true source positions by a white Gaussian noise with a small variance σ^2, representing calibration inaccuracies.

Our goal is to locate the source, associated with a new set of microphone measurements (1), based on a model inferred from the training set. For this purpose, we first extract appropriate features from the measured signals (1), which contain the relevant information regarding the varying source position. Then, we perform an interpolation of training positions with weights that are proportional to distances between the computed features. In Section 3, we discuss the feature extraction process, performed for each individual node. A unifying model, fusing the information inferred from the features of the different nodes, is presented in Section 4. Based on this model, a supervised Bayesian localizer is proposed in Section 5.

3. FEATURE EXTRACTION

The process of feature extraction is of great importance in machine learning and data mining algorithms [26]. The idea is to extract informative features from the *raw* data, which will best serve the subsequent learning task. Recently, we have presented several methods for source localization and tracking [20–23], all of which utilize features based on the relative transfer function (RTF), which is defined as the ratio between the acoustic transfer functions of two adjacent microphones [27]. The RTFs (concatenated for a certain frequency band) have a complex high-dimensional structure, representing the various reflections from the different surfaces defining the enclosure. However, when the acoustic conditions are approximately fixed, the variations of the RTFs are mainly attributed to the different source positions. Therefore, in such scenarios, the RTFs admit a compact representation on a low-dimensional *manifold*.

In this paper, we propose to use new feature vectors, utilizing the unique characteristics of unidirectional arrays. Directional microphones, pointed to different angles, perceive different parts of the reflections, depending on their radiation pattern and orientation. For example, consider an array with one element directed towards the source and another element oriented in the opposite direction. The front microphone receives both the direct signal and the reverberant part, whereas the opposite microphone receives only the reflections. Assuming that the reverberations can be modeled as a diffuse sound field, propogating incoherntly in all directions, they are preceived similarly by both elements. Based on this obsrvation, in [25], it was proposed to utilize this type of architecture for reverberated speech quality assessment. It was shown that the energy ratio between the direct microphone and the opposite microphone can be used to blindly estimate the direct-to-reverberant energy ratio (DRR). In addition, experimental results demonstrated that the computed ratio is monotonically varying with respect to the source distance from the microphone. Power ratios based on measurments of directional sensors are commonly used for bearing estimation in radar [28] and sonar [29] applications.

Motivated by the analysis and observations of [25], we define a new feature vector based on the power ratios between the microphones in each node. The power ratio between the ith and the jth microphones of the mth node is evaluated by:

$$R_{ij}^m(t,\mathbf{p}) = \frac{S_i^m(t,\mathbf{p})}{S_j^m(t,\mathbf{p})} \tag{2}$$

where the powers $S_i^m(t,\mathbf{p}),\ i \in \{1,\ldots,J\},\ m \in \{1,\ldots,M\}$ are computed over short time-periods of length T:

$$S_i^m(t,\mathbf{p}) = \frac{1}{T}\int_{t-T}^{t} (y_i^m(\tau,\mathbf{p}))^2 d\tau \tag{3}$$

where T is set to the range $20-40$ms to correspond to the speech quasi-stationarity time scale. Alternatively, the powers can be computed using an exponential averaging with parameter α:

$$S_i^m(t,\mathbf{p}) = (1-\alpha)S_i^m(t-T_s,\mathbf{p}) + \alpha(y_j^m(t,\mathbf{p}))^2. \tag{4}$$

where T_s is the sampling period. The value of α should be approximately set to T_s/T.

Let $R_{ij}^m(\mathbf{p})$ denote the average value of (2) over the observation interval. In the average ratio, we include only time periods with positive derivatives of the power of the first microphone, which are assumed to contain the direct part. Let $\mathcal{L}$ denote the set of L pairs of indices from $\{1,\ldots,J\}$, with $L=\binom{J}{2}$. We define the feature vector $\mathbf{r}^m$ of size $L\times 1$ as a concatenation of the set of power ratios of the mth node $\left\{10\log_{10}\left(R_{ij}^m(\mathbf{p})\right)\right\}_{(i,j)\in\mathcal{L}}$, computed in logarithmic scale.

In the training phase, we compute the features (2) for each training position $\bar{\mathbf{p}}_n$ and form the set $\{\bar{\mathbf{r}}_n^m\}_{m,n=1}^{M,N}$. In the test phase, we compute the feature vectors $\{\mathbf{r}_{\mathrm{t}}^m\}_{m=1}^M$ for each new arriving measurement from an unknown position $\mathbf{p}_{\mathrm{t}}$. For notational clarification, we emphasize that the training positions and features are marked by bars, whereas test positions and features are marked by subscript t.

Compared to the RTFs, which contain a large amount of parameters to be estimated, the ratio-based features are low-dimensional and efficient to compute. Moreover, we show in Section 6 that they are advantageous in acoustic scenarios with low reverberations, as well as in noisy scenarios with high levels of background noise. In fact, the ratio-based features are coarser compared to the detailed high-dimensional RTFs. Thus, they are more suitable to capture the variability of low reverberation acoustic scenarios, where the high-dimensional RTFs contain redundant information. In addition, such coarse features are preferable in noisy scenarios, where the estimation of the detained RTF-based features, consisting of a large number of parameters, yields large errors.

4. MULTI-VIEW GAUSSIAN PROCESS

Based on the defined features, we build a local model for each node describing a map of its set of features to the corresponding source positions. We show that the models of the different nodes represent different view points, and define a unifying mapping tying the different views together [22].

Assuming that the nodes are spatially distributed in the enclosure of interest, the features of each node have unique patterns, representing different perspectives on the acoustic scene. To build a local model, we construct a graph G^m for each node m. The graph vertices are defined by the training positions, and the edges connecting them are weighted according to distances between the corresponding features. The weights are computed using a pairwise *kernel* function. Here, a Laplacian (exponential) kernel is employed:

$$k^m\left(\bar{\mathbf{r}}_n^m,\bar{\mathbf{r}}_l^m\right) = \exp\left\{-\frac{\|\bar{\mathbf{r}}_n^m-\bar{\mathbf{r}}_l^m\|_2}{\varepsilon_m}\right\} \tag{5}$$

where $\|\cdot\|_2$ denotes the l_2 norm, and ε_m is a scaling parameter. Note that we use a Laplacian kernel rather than the typical Gaussian kernel, since it decays slower, hence more appropriate for the usage of simplified low-dimensional features, which are smoother and more regular with respect to the source position.

The relations induced by the kernel (5) represent the local model of the mth node. We would like to merge the different views presented by the nodes, and then relate the inferred relations among the features to the relations between the corresponding source positions.

We define a zero-mean Gaussian process $f^m \sim \mathcal{GP}(0, \tilde{k}^m)$ which attaches each feature $\bar{\mathbf{r}}_n^m$ with the corresponding source position $f_n^m \equiv f^m(\bar{\mathbf{r}}_n^m)$. Note that f^m is a scalar function representing either x, y or z coordinate of the source position. Since the same derivation applies to each coordinate separately, the coordinate subscript is omitted. The covariance of the process is defined by:

$$\text{cov}(f_n^m, f_l^m) \equiv \tilde{k}^m(\bar{\mathbf{r}}_n^m, \bar{\mathbf{r}}_l^m) = \sum_{w=1}^{N} k^m(\bar{\mathbf{r}}_n^m, \bar{\mathbf{r}}_w^m) k^m(\bar{\mathbf{r}}_l^m, \bar{\mathbf{r}}_w^m). \tag{6}$$

We should emphasize the difference between the kernels k and $\tilde{k}$, defined in (5) and in (6), respectively. While the kernel k defines proximity by measuring the pairwise relations between samples, the kernel $\tilde{k}$ also examines their relations to the given training samples. When two samples convey similar connections to the other training samples, it indicates that they themselves are closely related, and vice versa. The covariance between the processes of different nodes, m and g, is defined in a similar way by:

$$\text{cov}(f_n^m, f_l^g) = \sum_{w=1}^{N} k^m(\bar{\mathbf{r}}_n^m, \bar{\mathbf{r}}_w^m) k^g(\bar{\mathbf{r}}_l^g, \bar{\mathbf{r}}_w^g). \tag{7}$$

Here, the intra-relations within the mth and gth nodes are evaluated separately by the kernels k^m and k^g, respectively, and then they are composed together in (7).

The different perspectives introduced by the different nodes are fused by the Multi-View Gaussian Process (MVGP) f, defined as the average process of $\{f^m\}_{m=1}^M$:

$$f_n \equiv f(\bar{\mathbf{r}}_n) = \frac{1}{M}\left(f^1(\bar{\mathbf{r}}_n^1) + f^2(\bar{\mathbf{r}}_n^2) + \ldots + f^M(\bar{\mathbf{r}}_n^M)\right). \tag{8}$$

where $\bar{\mathbf{r}}_n = [(\bar{\mathbf{r}}_n^1)^T, \ldots, (\bar{\mathbf{r}}_n^M)^T]^T$. The MVGP is zero-mean and its covariance function is given by:

$$\begin{aligned}\text{cov}(f_n, f_l) &= \frac{1}{M^2}\text{cov}\left(\sum_{m=1}^{M} f_n^m, \sum_{g=1}^{M} f_l^g\right) \\ &= \frac{1}{M^2}\sum_{m,g=1}^{M} \text{cov}(f_n^m, f_l^g).\end{aligned} \tag{9}$$

Substituting (6) and (7) into (9) yields:

$$\begin{aligned}&\text{cov}(f_n, f_l) \equiv \tilde{k}(\bar{\mathbf{r}}_n, \bar{\mathbf{r}}_l) \\ &\quad = \frac{1}{M^2}\sum_{m,g=1}^{M}\sum_{w=1}^{N} k^m(\bar{\mathbf{r}}_n^m, \bar{\mathbf{r}}_w^m) k^g(\bar{\mathbf{r}}_l^g, \bar{\mathbf{r}}_w^g).\end{aligned} \tag{10}$$

The covariance of the process f aggregates the pairwise relations between each two nodes, enhancing observations common to pairs of nodes and ignoring inconsistent observations. It was shown in [30] that such covariance terms enhance the dominant parameters governing the data, which, in our case, correspond to the varying source positions.

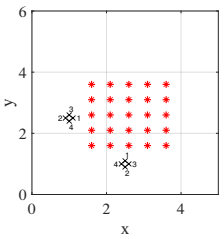


Fig. 1. Setup: The black 'x' marks denote the microphone positions, and the red '*' marks denote the training positions.

5. LOCALIZATION ALGORITHM

Based on the defined model we formulate a Bayesian localizer, applying a regression between the given training positions and the covariance terms of the MVGP.

In the test phase, we receive new measurements of a source located in an unknown position p_t. Here as well, we analyse each coordinate (x, y or z) separately, hence a scalar notation is used. Our goal is then to estimate the position of the source based on the extracted features $\mathbf{r}_\text{t} = [(\mathbf{r}_\text{t}^1)^T, \ldots, (\mathbf{r}_\text{t}^M)^T]^T$. The position of a new test point can be considered as a sample of the defined MVGP, i.e. $f_\text{t} = f(\mathbf{r}_\text{t})$. Therefore, the current position and all the training positions $\mathbf{p}_N = [\bar{p}_1, \ldots, \bar{p}_N]^T$ are jointly Gaussian, and their conditional distribution is given by:

$$\begin{bmatrix}\mathbf{p}_N \\ p_\text{t}\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}_{N+1}, \begin{bmatrix}\tilde{\mathbf{\Sigma}}_N + \sigma^2\mathbf{I}_N & \tilde{\mathbf{\Sigma}}_{N\text{t}} \\ \tilde{\mathbf{\Sigma}}_{N\text{t}}^T & \tilde{\Sigma}_\text{t}\end{bmatrix}\right) \tag{11}$$

where $\mathbf{I}_N$ is the unit matrix of size $N \times N$ and

$$\begin{aligned}(\tilde{\mathbf{\Sigma}}_N)_{nl} &= \tilde{k}(\bar{\mathbf{r}}_n, \bar{\mathbf{r}}_l),\ 1 \le l, n \le N \\ (\tilde{\mathbf{\Sigma}}_{N\text{t}})_n &= \tilde{k}(\bar{\mathbf{r}}_n, \mathbf{r}_\text{t}),\ 1 \le n \le N \\ \tilde{\Sigma}_\text{t} &= \tilde{k}(\mathbf{r}_\text{t}, \mathbf{r}_\text{t}).\end{aligned}$$

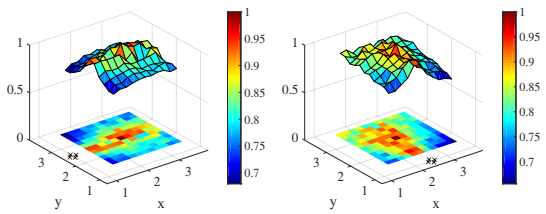
Accordingly, the minimum mean squared error (MMSE) estimator of p_t is given by its conditional mean, given all the training positions:

$$\hat{p}_\text{t} = \tilde{\mathbf{\Sigma}}_{N\text{t}}^T\left(\tilde{\mathbf{\Sigma}}_N + \sigma^2\mathbf{I}_N\right)^{-1}\mathbf{p}_N. \tag{12}$$

6. SIMULATION RESULTS

The ability of the proposed estimator to locate the source is examined in this section. We simulate a room of size $6 \times 5 \times 4$m with $M = 2$ nodes, each of which consists of $J = 4$ cardioid microphones, pointing in perpendicular directions. The AIRs are simulated using an efficient implementation [31] of the image method [32]. The source positions are assumed to be confined to a rectangular region of size 2×2m, with a fixed height of 2m. The room setup and the microphone constellation are illustrated in Fig. 1.

The training set is formed by a grid of $N = 25$ positions with a resolution of 0.5m between adjacent positions. We set the variance of the labelling errors to $\sigma^2 = 0.001$. The performance is averaged over 100 test positions in the designated region. The measured signals (both training and test), are formed by convoluting 2.5s long TIMIT sentences, with the corresponding simulated AIRs, with the addition of a white Gaussian noise. The sampling rate is set to $f_s = 16$kHz.



(a) First Node (b) Second Node

Fig. 2. Relations induced by the kernel defined over the features of a specific node. The colors correspond to the kernel values computed by (5) between the middle point of the rectangular region and all other training points. The 'x' marks denote the microphone positions of the associated node.

A demonstration of the local models formed by the two nodes is given in Fig. 2. The colors correspond to the kernel values computed by (5) between the middle training point of the rectangular region and all other training points. The 'x' marks denote the microphone positions of the corresponding node. It can be seen that indeed each node exhibits a unique spatial map.

We compare the localization results of the proposed estimator (12), with features defined based on power ratios (2), and with features based on RTF values [22]. The power ratios are computed by (2) with $T = 32$ms, and averaged across time to form the vectors $\mathbf{r}^1, \mathbf{r}^2 \in \mathbb{R}^L$, with $L = \binom{4}{2} = 6$. The RTFs consist of 100 frequency bins ($0 - 0.8$kHz), estimated using Welch's method with windows of 2048 samples and 75% overlap. The RTFs are computed between the front microphone $j = 1$ and each of the other microphones $j = \{2, 3, 4\}$ for each node, constituting a concatenated feature vectors with $3 \cdot 100 = 300$ elements each. As a reference, we also compare to nearest-neighbour estimator with ratio-based features. Comparisons with the RTF-based features to other methods were conducted in [22].

The root mean squared errors (RMSEs) of the algorithms, averaged over 100 trials, are presented in Fig. 3. Figure 3(a) depicts the results for different reverberation times ranging between 150ms and 700ms, with white Gaussian noise with signal to noise ratio (SNR) fixed to 20dB. Figure 3(b) illustrates the results for different SNR levels ranging between 0dB and 20dB, with reverberation time fixed to 300ms.

We observe that the nearest-neighbor estimator is inferior with respect to the proposed localizer in nearly every scenario (12). For high SNR conditions (20dB), the RTF-based and the ratio-based implementations yield comparable results, with a slight advantage for the ratio-based localizer in low reverberation, and for the RTF-based localizer in high reverberation. Regarding noise levels, it can be seen that the ratio-based localizer is more robust to noise, showing superiority for high noise levels.

The ratio-based features and the RTF-based features represent different acoustic fingerprints, both encoding the variations in the corresponding source positions. Comparing between the two, the ratio-based features represent a low-dimensional coarse version of the high-dimensional RTF-based features. Due to their simplified structure, the ratio-based features are specifically tailored to simple acoustic scenarios, characterized by low reverberations. Conversely, in these scenarios, the RTF-based features have a redundant representation. Due to their redundancy, the process of mapping the RTFs to source positions becomes unnecessarily intricate, leading to inferior localization results. Similarly, when noise level is high, feature estimation is less accurate, and therefore the coarse ratio-based features, which depend on fewer parameters, are more robust. However,

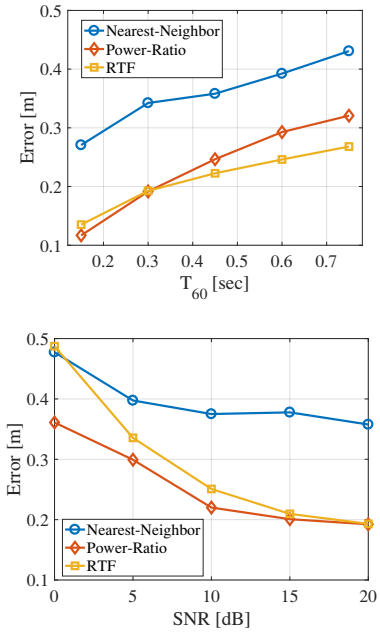


Fig. 3. RMSE obtained (a) for different reverberation times between 150ms to 700ms (SNR=20dB), and for different SNR levels between 0dB to 20dB ($T_{60} = 300$ms).

in high reverberations and high SNR conditions the RTF-based features better capture the complexity of the acoustic scenario, hence provide better localization results compared to the ratio-based features.

Regarding computational complexity, the ratio-based features are much simpler to compute compared to the RTF-based features. In this simulation, using a Matlab implementation on a standard PC (CPU Intel Core2 Quad 3.7 GHz, RAM 8 GB), the computation time of the ratio-based features and of the RTF-based features for a single point takes on average 7.5ms and 197ms, respectively. We conclude that the proposed ratio-based features are more efficient and are advantageous in low reverberations or low SNR conditions.

7. CONCLUSIONS

A robust source localization algorithm is derived based on computationally efficient features. The new features are based on power ratios evaluated between several unidirectional sensors in an array of microphones. It is shown that the features computed for each array reveal unique patterns, representing different view points. Merging the different perspectives of several arrays is carried out by defining a MVGP, whose covariance function aggregates relations observed by pairs of nodes. The derived Bayesian estimator performs a regression over training positions of prerecorded measurements, weighted by the covariance terms of the MVGP. The algorithm efficiency and robustness is demonstrated under noisy and reverberation conditions. Localization results demonstrate the advantage of the proposed features in low reverberation acoustic environments or in the presence of high background noises. These results indicate that the proposed supervised localization scheme can be generally applied to different type of features, extracted from the raw data. The different features represent unique acoustic fingerprints, encoding the corresponding source positions. The stage of feature extraction is of great importance to the localization results, and should be determined in correspondence to the environmental and acoustical conditions.

8. REFERENCES

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

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

[3] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer Berlin Heidelberg, 2001, pp. 157–180.

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

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

[7] T. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

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

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

[10] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[11] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

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

[13] T. D. Abhayapala and H. Bhatta, "Coherent broadband source localization by modal space processing," in *Proc. 10th international conference on telecommunications*, vol. 2, 2003, pp. 1617–1623.

[14] J. Le Roux, P. T. Boufounos, K. Kang, and J. R. Hershey, "Source localization in reverberant environments using sparse optimization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 4310–4314.

[15] A. Asaei, H. Bourlard, M. J. Taghizadeh, and V. Cevher, "Model-based sparse component analysis for reverberant speech localization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1439–1443.

[16] 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, 2011.

[17] 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.

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

[19] N. Bertin, S. Kitić, and R. Gribonval, "Joint estimation of sound source location and boundary impedance with physics-driven cosparse regularization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6340–6344.

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

[21] ——, "Semi-supervised sound source localization based on manifold regularization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1393–1407, 2016.

[22] ——, "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, 2017.

[23] ——, "Speaker tracking on multiple-manifolds with distributed microphones," in *Proc. Interntional Conference on Latent Variable Analysis and Signal Seperation (LVA/ICA)*, 2017, pp. 59–67.

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

[25] R. Berkun and I. Cohen, "Microphone array power ratio for quality assessment of reverberated speech," *EURASIP journal on advances in signal processing*, vol. 2015, no. 1, p. 49, 2015.

[26] I. Guyon and A. Elisseeff, "An introduction to feature extraction," in *Feature Extraction: Foundations and Applications*, I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds. Springer Berlin Heidelberg, 2006, pp. 1–25.

[27] 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, Aug. 2001.

[28] S. M. Sherman and D. K. Barton, *Monopulse principles and techniques*. Artech House, 2011.

[29] B. Maranda, "The statistical accuracy of an arctangent bearing estimator," in *Proc. OCEANS*, vol. 4, 2003, pp. 2127–2132.

[30] R. R. Lederman and R. Talmon, "Learning the geometry of common latent variables using alternating-diffusion," *Applied and Computational Harmonic Analysis*, 2015.

[31] E. A. P. Habets, "Room impulse response (RIR) generator," https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, Jul. 2006.

[32] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.