Acoustic Source Position Estimation based on Multi-Feature Gaussian Processes

Andreas Brendel, Ingo Altmann and Walter Kellermann
Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg,
Cauerstr. 7, D-91058 Erlangen, Germany, Andreas.Brendel@FAU.de

***Abstract*—Gaussian Processes, representing a Bayesian framework for regression, were already previously shown to allow effective range estimation in highly reverberant and noisy scenarios from a single pair of microphones when using the Coherent-to-Diffuse Power Ratio as a feature. In this work we investigate how Gaussian Process regression can jointly estimate range and Direction of Arrival by using the Coherent-to-Diffuse Power Ratio and an additional Direction of Arrival estimation feature (e.g., MUSIC) to achieve an estimate of the source position, based on a single concentrated array requiring only two sensors as a minimum.**

***Index Terms*—Gaussian process regression, acoustic source localization**

I. Introduction

The estimation of the position of an acoustic source is an important task for many signal processing applications as this positional information allows to, e.g., steer a camera to the source of interest [1] or to control speech enhancement algorithms [2]. In principle, position estimation may be achieved by estimating the Direction of Arrival (DOA) and the distance of the source under consideration. A significant amount of research has been dedicated to the task of DOA estimation in the last decades: Subspace methods including the well-known MUltiple SIgnal Classification (MUSIC) algorithm [3] and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) algorithm [4] have been introduced. Steered Power Response (SRP)-based methods have proven to be powerful in reverberant scenarios [5]. Blind Source Separation (BSS) demixing filters have been exploited to extract DOA information [6] and Expectation Maximization (EM)-based DOA estimators have been proposed [7]. However, many DOA estimation algorithms using linear microphone arrangements suffer from systematic estimation errors, especially for high reverberation times [8]. To address this issue, learning-based DOA estimation methods have been proposed, e.g., based on neural networks [9] or manifold learning [10].

On the contrary, distance estimation of an acoustic source is much less investigated. If no precise knowledge about the physical room parameters is available, the distance estimation methods usually rely on machine learning techniques [11], [12]. Following this direction, the Coherent-to-Diffuse power Ratio (CDR) has been used in [13] for learning-based distance estimation. The proposed method has been developed further for distributed training [14], distributed position estimation [15] and fixed budget learning [16] in acoustic sensor networks.

In many cases, not only the DOA or the distance, but the actual position of the source is of interest. A classification method for a learned set of source positions has been proposed in [17]. Several methods for acoustic source localization have been proposed relying on spatially distributed sensor nodes: Manifold learning has been applied to estimate the source position in a previously trained acoustic enclosure in [18]. An EM-based method for localization and tracking of multiple simultaneously active sources relying on pairwise relative phase ratios has been proposed in [19]. A general scheme for triangulation of multiple DOA estimates has been developed in [20].

In this contribution, we build upon previous publications on CDR-based distance estimation and aim at a learning-based acoustic position estimation algorithm relying on a single concentrated Uniform Linear Array (ULA) without the need of observations from distributed microphone arrays state-of-the-art methods rely on. To this end, we extend the current CDR-based approaches by a directional feature and we show that DOA estimation benefits from a training phase for bias removal and the integration of knowledge about the source distance. Reversely, the incorporation of the DOA knowledge is shown to assist the distance estimation. The proposed method uses Gaussian Process Regression (GPR) [21] as a machine learning tool to learn the mapping of the features to the position of the source.

II. Signal Model and Feature Extraction

We consider an ULA consisting of M microphones with microphone spacing d_{mic}, capturing a single acoustic wideband source. The Short-Time Fourier Transform (STFT) domain representation of the microphone channels, stacked in the vector

$$\mathbf{x}(t,k) = [x_1(t,k), \ldots, x_M(t,k)]^{\mathrm{T}} \tag{1}$$

with time index t and frequency index k, can be expressed as

$$\mathbf{x}(t,k) = \mathbf{h}(\theta,k)s(t,k) + \mathbf{n}(t,k). \tag{2}$$

Here, $s(t,k)$ denotes the clean source signal and

$$\mathbf{n}(t,k) = [n_1(t,k), \ldots, n_M(t,k)]^{\mathrm{T}} \tag{3}$$

This work was supported by DFG under contract no <Ke890/10-1> within the Research Unit FOR2457 "Acoustic Sensor Networks"

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

the received signal components corresponding to reverberation and sensor noise. The direct path propagation of the source is modeled by the source direction vector

$$\mathbf{h}(\theta,k)=[1,e^{-j2\pi f_k\tau(\theta)},\dots,e^{-j2\pi f_k(M-1)\tau(\theta)}]^{\mathrm{T}}, \quad (4)$$

which is dependent on the source DOA θ. Hereby, f_k denotes the physical frequency corresponding to frequency bin k and τ the time difference of arrival.

In the following, we describe a bin-wise directional and a bin-wise distance-related feature, which are transformed into broadband features through appropriate averaging operations. Note that these features may be exchanged if other features are more suitable for the task at hand.

A. Distance Feature

The employed distance feature is based on the CDR [22], which has been successfully used for dereverberation [22] and distance estimation [13]. Here, we exploit the fact that the reverberant sound energy is approximately constant with respect to the distance, whereas the energy of the coherent signal components is decaying with increasing distance. To estimate the CDR, the cross-Power Spectral Density (PSD) and the auto-PSDs of microphone pair (i,j) are estimated by recursive averaging

$$\hat{\Phi}_{x_ix_j}(t,k)=\lambda\hat{\Phi}_{x_ix_j}(t-1,k)+(1-\lambda)x_i(t,k)x_j^*(t,k), \quad (5)$$

where $i,j\in\{1,\dots,M\}$, $(\cdot)^*$ denotes complex conjugation and $\lambda\in[0,1]$ is a forgetting factor. An estimate of the complex-valued spatial coherence of microphone signal pair (i,j) can be computed by

$$\hat{\Gamma}_x^{ij}(t,k)=\frac{\hat{\Phi}_{x_ix_j}(t,k)}{\sqrt{\hat{\Phi}_{x_ix_i}(t,k)\hat{\Phi}_{x_jx_j}(t,k)}}. \quad (6)$$

To avoid any influence of the directional feature on the CDR-based feature, we use the DOA-independent CDR estimator (7), proposed in [22], here. The estimator relies on a model for the coherence of a diffuse soundfield

$$\Gamma_n^{ij}(k)=\frac{\sin(2\pi f_k(j-i)d_{\mathrm{mic}}/c)}{2\pi f_k(j-i)d_{\mathrm{mic}}/c}, \quad (8)$$

where c denotes the speed of sound and $i<j$. To finally obtain a scalar broadband feature, we average over the observation time interval containing T samples and the considered frequency range $[f_{k_{\min}},f_{k_{\max}}]$

$$\gamma_{\mathrm{r}}=\frac{1}{T(k_{\max}-k_{\min}+1)}\sum_{t=1}^{T}\sum_{k=k_{\min}}^{k_{\max}}\frac{1}{\widehat{\mathrm{CDR}}^{ij}(t,k)+1}. \quad (9)$$

By construction, $\gamma_{\mathrm{r}}\in\mathbb{W}_{\gamma_{\mathrm{r}}}=[0,1]$ holds. Note that the CDR is only defined for a pair of microphones.

B. Directional Feature

For the directional feature, we use the well-known MUSIC algorithm [3]. MUSIC is a subspace method which relies on the eigenvalue decomposition of the cross power spectral density matrix $\mathbf{\Phi_{xx}}(k)$. The signal subspace is of dimension one and the eigenvectors spanning the noise subspace are computed by eigenvalue decomposition and are stacked in the matrix

$$\mathbf{V}_n(k)=[\mathbf{v}_2(k),\dots,\mathbf{v}_M(k)]. \quad (10)$$

This matrix is employed to calculate the MUSIC pseudo spectrum

$$P_{\mathrm{MU}}(\theta,k)=\frac{1}{\mathbf{h}^{\mathrm{H}}(\theta_c,k)\mathbf{V}_n(k)\mathbf{V}_n^{\mathrm{H}}(k)\mathbf{h}(\theta_c,k)+\delta}, \quad (11)$$

where $\delta>0$ is a regularization term to avoid division by zero. The directional scalar broadband feature is calculated by averaging the MUSIC pseudo spectrum over the considered frequency range $[f_{k_{\min}},f_{k_{\max}}]$ and maximization

$$\gamma_\phi=\underset{\theta_c\in\mathcal{C}}{\operatorname{argmax}}\left(\frac{1}{k_{\max}-k_{\min}+1}\sum_{k=k_{\min}}^{k_{\max}}(P_{\mathrm{MU}}(\theta,k))^{-1}\right)^{-1}. \quad (12)$$

Here, $f_{k_{\min}}$ and $f_{k_{\max}}$ denote the physical frequency corresponding to the minimum and maximum frequency index $k_{\min}$ and $k_{\max}$, respectively. The grid of candidate target directions is denoted as

$$\mathcal{C}=\{\theta_\nu\in[-90^\circ,90^\circ]|\theta_\nu=-90^\circ+\nu\Delta\theta,\nu\in\mathbb{N}_0\}, \quad (13)$$

with angular resolution $\Delta\theta$. Note that the feature values are bounded by $\gamma_\phi\in\mathbb{W}_{\gamma_\phi}=[-90^\circ,90^\circ]$.

III. Learning-based Position Estimation

In the following, we develop the proposed learning-based position estimation algorithm. As we want to account for different characteristic properties for the DOA label ζ_ϕ and the distance label ζ_{r}, we derive a generic regression model that can be adapted to two different regression functions, which suit both labels individually. To encode that, we use the variable $\zeta\in\{\zeta_{\mathrm{r}},\zeta_\phi\}$ for both labels in the following derivation.

We model the label ζ to be related with the two-dimensional feature

$$\boldsymbol{\xi}=\begin{bmatrix}\gamma_{\mathrm{r}} & \gamma_\phi\end{bmatrix}^{\mathrm{T}}\in\mathbb{D}=\mathbb{W}_{\gamma_{\mathrm{r}}}\times\mathbb{W}_{\gamma_\phi} \quad (14)$$

via the unknown function f_ζ by

$$\zeta=f_\zeta(\boldsymbol{\xi})+\epsilon \quad \text{with} \quad \epsilon\sim\mathcal{N}(0,\sigma_\epsilon^2). \quad (15)$$

Since the functions' family is unknown a priori, we rely on GPR for learning the functional relationships between feature vector $\boldsymbol{\xi}$ and both labels $\zeta\in\{\zeta_{\mathrm{r}},\zeta_\phi\}$. The latent function f_ζ is modeled to follow a Gaussian Process (GP)

$$f_\zeta\sim\mathcal{GP}(m_\zeta,k_\zeta), \quad (16)$$

with mean function $m_\zeta:\mathbb{D}\to\mathbb{R}_+$ and covariance function $k_\zeta:\mathbb{D}\times\mathbb{D}\to\mathbb{R}_+$. The hyperparameters of the mean and covariance function can be calculated in a learning phase, e.g.,

$$\widehat{\mathrm{CDR}}^{ij} = \frac{1}{\left|\hat{\Gamma}_x^{ij}\right|^2 - 1}\left(\Gamma_n^{ij}\ \mathrm{Re}\left\{\hat{\Gamma}_x^{ij}\right\} - \left|\hat{\Gamma}_x^{ij}\right|^2 - \sqrt{\left(\Gamma_n^{ij}\right)^2\ \mathrm{Re}\left\{\hat{\Gamma}_x^{ij}\right\}^2 - \left(\Gamma_n^{ij}\right)^2\ \left|\hat{\Gamma}_x^{ij}\right|^2 + \left(\Gamma_n^{ij}\right)^2 - 2\,\Gamma_n^{ij}\ \mathrm{Re}\left\{\hat{\Gamma}_x^{ij}\right\} + \left|\hat{\Gamma}_x^{ij}\right|^2}\right) \quad (7)$$

by optimizing the marginal likelihood [21]. In the following, features and labels of the learning phase are marked by a tilde $\tilde{(\cdot)}$. The features used in the learning phase are collected in the set

$$\mathcal{F} = \left\{\tilde{\boldsymbol{\xi}}_n | n = 1, \ldots, N_{\text{train}}\right\}, \quad (17)$$

where N_{train} denotes the number of training data points. Based on the collected feature values and the chosen covariance function, the kernel matrix

$$\mathbf{K}_\zeta = [k_\zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)]_{i,j} \text{ with } \boldsymbol{\xi}_i, \boldsymbol{\xi}_j \in \mathcal{F} \text{ and } i, j = 1, \ldots, N_{\text{train}} \quad (18)$$

can be constructed. Similarly, the covariance vector of the features of the training set with an unlabeled feature $\boldsymbol{\xi}$ is defined as

$$\mathbf{k}_\zeta(\boldsymbol{\xi}) = [k_\zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi})]_i \text{ with } \boldsymbol{\xi}_i \in \mathcal{F} \text{ and } i = 1, \ldots, N_{\text{train}} \quad (19)$$

and the vector of mean function values is expressed as

$$\mathbf{m}_\zeta = [m_\zeta(\tilde{\boldsymbol{\xi}}_1), \ldots, m_\zeta(\tilde{\boldsymbol{\xi}}_{N_{\text{train}}})]^{\mathrm{T}}, \quad (20)$$

where m_ζ will be defined later. We model the function value $f_\zeta(\boldsymbol{\xi})$ of an unlabeled feature $\boldsymbol{\xi}$ and the labels for the direction or distance of the training data points

$$\tilde{\boldsymbol{\zeta}} = [\tilde{\zeta}_1, \ldots, \tilde{\zeta}_{N_{\text{train}}}]^{\mathrm{T}} \quad \text{with} \quad \tilde{\zeta}_i \in \{\tilde{\zeta}_{\mathrm{r,i}}, \tilde{\zeta}_{\phi,\mathrm{i}}\} \quad (21)$$

to be jointly Gaussian distributed

$$\begin{bmatrix}\tilde{\boldsymbol{\zeta}} \\ f_\zeta(\boldsymbol{\xi})\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}_\zeta \\ m_\zeta(\boldsymbol{\xi})\end{bmatrix}, \begin{bmatrix}\mathbf{K}_\zeta + \sigma_\epsilon^2\mathbf{I} & \mathbf{k}_\zeta(\boldsymbol{\xi}) \\ \mathbf{k}_\zeta^{\mathrm{T}}(\boldsymbol{\xi}) & k_\zeta(\boldsymbol{\xi}, \boldsymbol{\xi})\end{bmatrix}\right). \quad (22)$$

Now, we can calculate the mean of the posterior distribution of the unknown label, i.e., the noiseless function value $f_\zeta(\boldsymbol{\xi})$, which constitutes the desired regression function for label $\zeta \in \{\zeta_\mathrm{r}, \zeta_\phi\}$ [21]

$$\hat{\zeta} = f_\zeta(\boldsymbol{\xi}) = m_\zeta(\boldsymbol{\xi}) + \mathbf{k}_\zeta^{\mathrm{T}}(\boldsymbol{\xi})\left(\mathbf{K}_\zeta + \sigma_\epsilon^2\mathbf{I}\right)^{-1}\left(\tilde{\boldsymbol{\zeta}} - \mathbf{m}_\zeta\right). \quad (23)$$

In the following, we will discuss the choices for the mean and covariance functions used to specify the individual GP models linked to the DOA ζ_ϕ and the distance ζ_r.

A. DOA Estimation

DOA estimation is usually done by using simple geometric and physical models which allow to obtain an estimate without the need of a learning phase. However, for many algorithms, DOA estimation with ULAs suffers from systematic errors, especially for large distances and high reverberation times [8]. Hence, a learning phase is optional but beneficial and, for Time Difference of Arrival (TDOA)-based methods, does not need different geometric models for near-field and far-field conditions.

Since DOA estimation algorithms based on geometric models already yield estimates for ζ_ϕ, we incorporate these results by choosing a linear mean function[1]

$$m_{\zeta_\phi}(\boldsymbol{\xi}) = \beta_\mathrm{r}\gamma_\mathrm{r} + \beta_\phi\gamma_\phi, \quad (24)$$

where β_r and β_ϕ denote coefficients controlling the slope of the linear contributions. For the covariance function, we choose the well-known Gaussian kernel

$$k_{\zeta_\phi}(\boldsymbol{\xi}, \boldsymbol{\xi}') = \alpha_\phi \exp\left(-\frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}')^{\mathrm{T}}\mathbf{L}_\phi^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}')\right), \quad (25)$$

where α_ϕ, the so-called signal variance, controls the allowed deviation from the mean function. The smoothness of the latent function is controlled by two different length scale parameters l_r and l_ϕ for the distance and the directional feature, respectively, which are contained in the matrix

$$\mathbf{L}_\phi = \mathrm{diag}\left\{l_\mathrm{r}^2, l_\phi^2\right\}. \quad (26)$$

Hereby, $\mathrm{diag}\{\cdots\}$ denotes the diagonal matrix with its arguments as entries on the main diagonal.

B. Distance Estimation

The relation of the chosen distance-related feature γ_r to the actual distance between microphone array and source depends on several room-specific physical parameters, which are unknown in practice. Hence, this relation has to be learned from observed data, i.e., learning is mandatory here. As we cannot assume reliable prior knowledge about the latent function, we take the common choice

$$m_{\zeta_\mathrm{r}}(\boldsymbol{\xi}) = 0 \quad \forall\boldsymbol{\xi} \in \mathbb{D} \quad (27)$$

for the mean function. For the covariance function, we choose a sum of two Gaussian kernels, one corresponding to a smooth behavior (s) of the latent function and one corresponding to a more dynamic behavior (d), to account for the smooth behavior of the regression function for a large range of distances and the strong slope for large distances (see [15])

$$k_{\zeta_\mathrm{r}}(\boldsymbol{\xi}, \boldsymbol{\xi}') = \sum_{i \in \{(\mathrm{s}),(\mathrm{d})\}} \alpha_\mathrm{r}^i \ldots \quad (28)$$
$$\ldots \exp\left(-\frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}')^{\mathrm{T}}\left(\mathbf{L}_\mathrm{r}^i\right)^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}')\right).$$

Hereby, the length scale parameters l_r^i and l_ϕ^i are included in the matrices

$$\mathbf{L}_\phi^i = \mathrm{diag}\left\{(l_\mathrm{r}^i)^2, (l_\phi^i)^2\right\} \quad \text{with} \quad i \in \{(\mathrm{l}), (\mathrm{s})\}. \quad (29)$$

[1]The linear term $\beta_\mathrm{r}\gamma_\mathrm{r}$ is motivated by a requirement of [23].

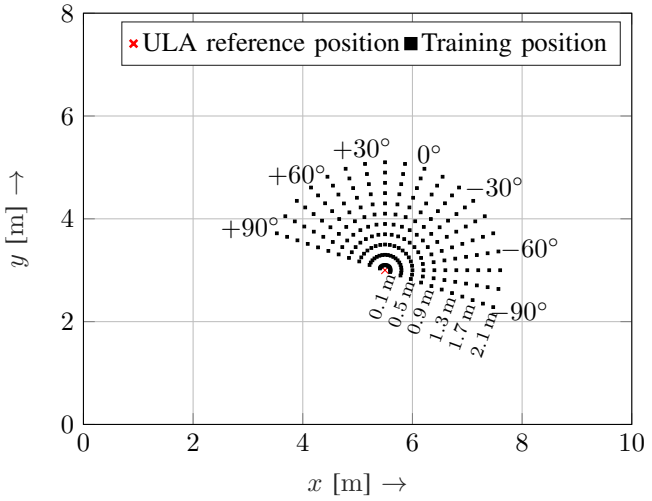

Fig. 1. Simulated room environment of dimensions $10\,\mathrm{m} \times 8\,\mathrm{m} \times 3.5\,\mathrm{m}$. The training positions used for creating the training data set are marked by black squares. The red cross marks the position of the arrays reference position.

$l_\mathrm{r},\ l_\mathrm{r}^{(s)}, l_\mathrm{r}^{(d)}$	$l_\phi, l_\phi^{(s)}, l_\phi^{(d)}$	$\beta_\mathrm{r}, \beta_\phi$	$\alpha_\phi, \alpha_\mathrm{r}^{(s)}, \alpha_\mathrm{r}^{(d)}$	σ_ϵ^2
1, 0, 0.15	0.5, 20, 0.25	0, 1	0.5, 0.5, 10	0.01

TABLE I
INITIAL VALUES FOR HYPERPARAMETER OPTIMIZATION

C. Position Estimation

For both regression functions, the hyperparameters $l_\mathrm{r}, l_\mathrm{r}^{(s)}, l_\mathrm{r}^{(d)}, l_\phi, l_\phi^{(s)}, l_\phi^{(d)}, \beta_\mathrm{r}, \beta_\phi, \alpha_\phi, \alpha_\mathrm{r}^{(s)}, \alpha_\mathrm{r}^{(d)}$ and σ_ϵ^2 are found by optimization of the marginal likelihood function [21] [23] before calculating the posterior mean function. Finally, the evaluation of both posterior mean functions yield the estimated source position in polar coordinates.

IV. Simulation Study

To assess the performance of the proposed algorithm, we conducted experiments in a simulated enclosure of dimensions $10\,\mathrm{m} \times 8\,\mathrm{m} \times 3.5\,\mathrm{m}$. All sources and microphones are placed at the same height of $1.5\,\mathrm{m}$, i.e., we restrict the localization to two dimensions. The Region of Interest (ROI) is chosen to be a half annulus with inner radius $0.1\,\mathrm{m}$ and outer radius $2.1\,\mathrm{m}$ centered at the ULA reference point. The ROI is covered by training positions with angular resolution of $10°$ and radial resolution of $0.2\,\mathrm{m}$, yielding a total of $N_\mathrm{train} = 209$ labeled training pairs, see Figure 1. Room Impulse Responses (RIRs) are simulated using the image source method [24] and the RIR generator [25] for various T_{60}. To simulate $M = 4$ microphone signals, the simulated RIRs are convolved with speech signals of $5\,\mathrm{sec}$ duration at a sampling frequency of $f_s = 16\,\mathrm{kHz}$ and white Gaussian noise is added for a specific Signal to Noise Ratio (SNR). Both spatial features are calculated in the frequency domain: The MUSIC feature is computed using non-overlapping rectangular windows of length $50\,\mathrm{ms}$ and the CDR-based feature with a von Hann window of $25\,\mathrm{ms}$ length and $12.5\,\mathrm{ms}$ frame shift.

For the MUSIC algorithm, all microphones, with a spacing of $d_\mathrm{mic} = 5\,\mathrm{cm}$ are used, whereas for the estimation of the CDR, only the outermost microphones are exploited. The considered frequency interval was chosen to be $[f_{k_\mathrm{min}}, f_{k_\mathrm{max}}] = [300\,\mathrm{Hz}, 4000\,\mathrm{Hz}]$ for both features. The resolution of the candidate target directions for the MUSIC algorithm was set to $\Delta\theta = 0.5°$. For the smoothing parameter, $\lambda = 0.9$ was chosen to alleviate influence of speech pauses. The optimization of the hyperparameters of the GPR models has been initialized with the values given in Table I.

To assess the performance of the proposed algorithm, $N_\mathrm{test} = 800$ test positions, regularly uniformly distributed over distance and directions within the ROI, are evaluated using different speech signals for training and test and the position error has been averaged over all N_test positions

$$e = \frac{1}{N_\mathrm{test}} \sum_{j=1}^{N_\mathrm{test}} \|\hat{\mathbf{p}}_j - \mathbf{p}_j\|_2. \tag{30}$$

Hereby, $\mathbf{p}_j$ denotes the jth test position and $\hat{\mathbf{p}}_j$ the corresponding estimate.

Typical trained regression functions for DOA as well as for distance estimation, i.e., posterior mean functions, are shown in Figure 2 for $T_{60} = 0.6\,\mathrm{s}$ and $\mathrm{SNR} = 20\,\mathrm{dB}$. The figure shows isolines, i.e., the set of feature vectors $\boldsymbol{\xi}$ which yield the same label estimate. It can be clearly seen that the regression function for the DOA varies for different distance features ζ_r, i.e., the systematic error of the DOA estimator, corrected by the regression function, depends on the estimated CDR and thus on the radial distance. Especially for endfire directions, the DOA is underestimated, which becomes more pronounced for large distances, reflected by large values for ζ_r. On the other hand, the distance-related feature shows a dependency on the direction of the source, which is pronounced between $-40°$ and $40°$. As the isolines are not circular or radial, respectively, it can be concluded that the two-dimensional feature vector containing a directional feature as well as a distance-related feature is beneficial for the regression task as directional information can assist distance estimation and vice versa. Figure 3 shows the average position error e for reverberation times $T_{60} \in \{0.4\,\mathrm{s}, 0.6\,\mathrm{s}, 0.8\,\mathrm{s}, 1\,\mathrm{s}\}$ and SNR values between $-10\,\mathrm{dB}$ and $30\,\mathrm{dB}$. The values show that the algorithm works robustly for a broad range of reverberation times and noise levels. The algorithm delivers precise position estimates almost independent of the SNR down to $\mathrm{SNR} = 0\,\mathrm{dB}$ before the performance degrades significantly.

V. Conclusions

In this contribution, we proposed a learning-based position estimation algorithm for a single concentrated ULA based on a directional and a distance-related feature. To this end, separate regression functions for DOA and distance of an acoustic source have been learned. Here, the directional feature was shown to be beneficial for the distance estimation task and vice versa. The next steps in the development of this method will include the extension of the algorithm to semi-supervised learning techniques. The generalization of the algorithm to sensors embedded into scatterers and the extension to multiple sources are also important next steps.

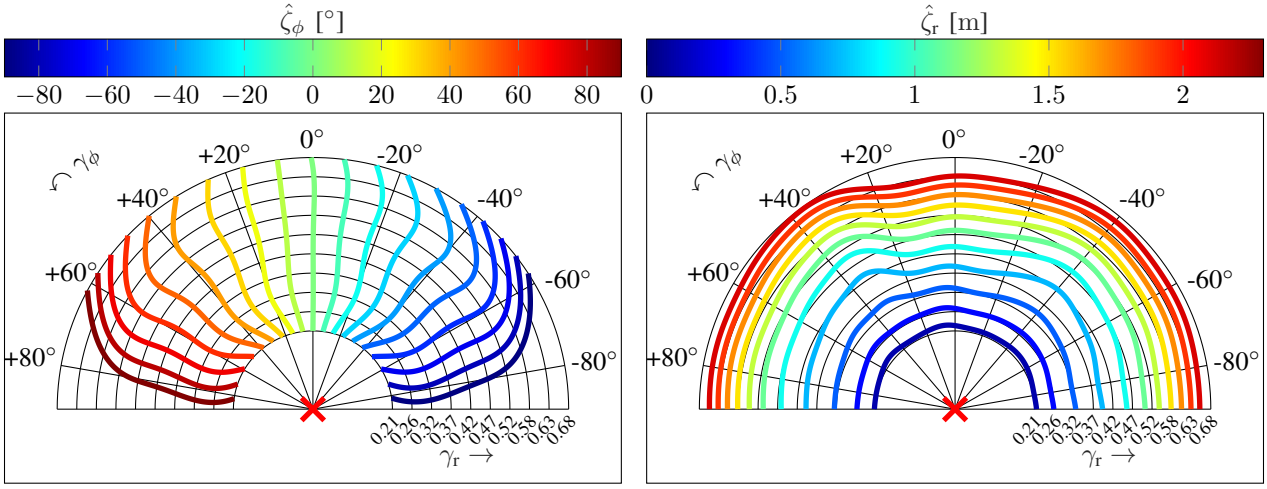

Fig. 2. Contour plot of the regression function for DOA ζ_ϕ (left panel) and distance ζ_r (right panel) for $T_{60} = 0.6\,\mathrm{s}$ and $\mathrm{SNR} = 20\,\mathrm{dB}$. The colored lines represent isolines of the underlying regression function, drawn in dependence of the underlying true labels.

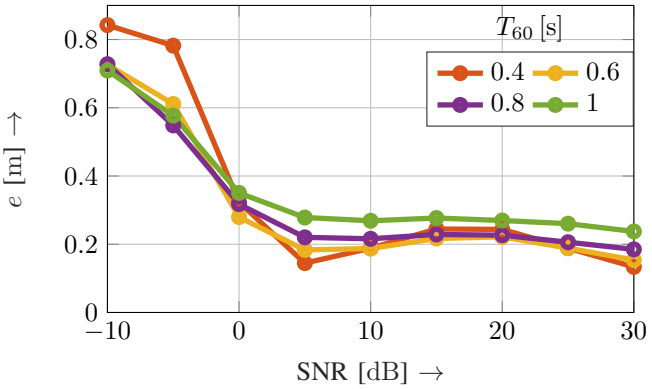

Fig. 3. Average position error e for different noise levels and reverberation times.

References

[1] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Munich, Germany, Oct. 1997.

[2] M. Taseska and E. A. P. Habets, "Informed Spatial Filtering for Sound Extraction Using Distributed Microphone Arrays," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 22, no. 7, pp. 1195–1207, Jul. 2014.

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

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

[5] J. H. DiBiase, "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays," PHD Thesis, Brown University, Providence, Rhode Island, May 2000.

[6] A. Lombard et al., "TDOA Estimation for Multiple Sound Sources in Noisy and Reverberant Environments Using Broadband Independent Component Analysis," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[7] Y. Dorfan et al., "Multiple DOA estimation and blind source separation using estimation-maximization," Eilat, Israel, Nov. 2016, pp. 1–5.

[8] F. Jacob and R. Haeb-Umbach, "On the Bias of Direction of Arrival Estimation Using Linear Microphone Arrays," in *12. ITG Symp. Speech Commun.*, Oct. 2016.

[9] X. Xiao et al., "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 2814–2818.

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

[11] S. Vesa, "Binaural Sound Source Distance Learning in Rooms," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 17, no. 8, pp. 1498–1507, Nov. 2009.

[12] Y. Lu and M. Cooke, "Binaural Estimation of Sound Source Distance via the Direct-to-Reverberant Energy Ratio for Static and Moving Sources," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1793–1805, Sep. 2010.

[13] A. Brendel and W. Kellermann, "Learning-Based Acoustic Source-Microphone Distance Estimation Using the Coherent-to-Diffuse Power Ratio," in *IEEE Int. Conf. on Acoustic, Speech and Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018.

[14] ——, "Distance Estimation of Acoustic Sources using the Coherent-to-Diffuse Power Ratio Based on Distributed Training," in *IEEE Int. Workshop on Acoustic Signal Enhancement*, Tokyo, Japan, Sep. 2018.

[15] ——, "Learning-based Acoustic Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio," in *European Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018.

[16] ——, "Distributed Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[17] P. Smaragdis and P. Boufounos, "Position and Trajectory Learning for Microphone Arrays," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 15, no. 1, pp. 358–368, Jan. 2007.

[18] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-Supervised Source Localization on Multiple Manifolds With Distributed Microphones," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 25, no. 7, pp. 1477–1491, Jul. 2017.

[19] O. Schwartz and S. Gannot, "Speaker Tracking Using Recursive EM Algorithms," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 22, no. 2, pp. 392–402, Feb. 2014.

[20] A. Griffin et al., "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 54–67, Feb. 2015.

[21] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, ser. Adaptive computation and machine learning. Cambridge, Mass: MIT Press, 2006.

[22] A. Schwarz and W. Kellermann, "Coherent-to-Diffuse Power Ratio Estimation for Dereverberation," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 23, no. 6, pp. 1006–1018, Jun. 2015.

[23] C. E. Rasmussen and H. Nickisch, "Gaussian Process Regression and Classification Toolbox." [Online]. Available: www.gaussianprocess.org/gpml/code/matlab/doc

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

[25] E. A. P. Habets, "Room Impulse Response Generator," Int. Audio Laboratories, Tech. Rep., Sep. 2010.