PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Probabilistic modeling for learning-based distance estimation

Andreas BRENDEL, Andy REGENSKY, and Walter KELLERMANN

Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstr. 7, D-91058 Erlangen, Germany, {Andreas.Brendel}@FAU.de.

Abstract

Estimating the position of an acoustic source is an essential step for many signal processing applications. Hence, many approaches for acoustic source localization have been proposed in recent years, whereby most of them are based on the estimation of the direction of arrival of a wavefront emitted by the acoustic source relative to an observing microphone array. In contrast, the estimation of the source-microphone distance is much less investigated. Especially for complex sound fields in acoustic enclosures, a promising approach is to apply machine learning approaches to learn the mapping of a distance-related feature to the corresponding source-microphone distance. Here, we focus on a recently proposed method for distance estimation based on the coherent-to-diffuse power ratio and Gaussian process regression. We investigate the influence of different probabilistic models reflected by the choice of kernel and mean functions of the Gaussian process employed for regression. The influence of these choices is quantified by experiments.

Keywords: Distance estimation, Gaussian process regression, Coherent-to-Diffuse power Ratio

1 INTRODUCTION

Knowledge about the position of an acoustic source is crucial for many signal processing applications. Therefore, significant research effort has been dedicated to the estimation of the Direction of Arrival (DOA) of acoustic sources, which is usually based on simplified geometric models [10, 14, 5].

On the other side, the estimation of the source-microphone distance is much less investigated. Two main approaches for acoustic distance estimation can be distinguished: Approaches relying on precise knowledge of the acoustic environment [12, 9] and learning-based approaches [16, 6, 7]. In this contribution, we consider a recently developed method of the second kind: The ratio of the powers of coherent and diffuse signal components has been proposed as a distance-related feature in [3]. Based on this, a localization strategy for acoustic sensor networks, where the training of a Gaussian Process Regression (GPR) model and position estimation is realized in a distributed manner, has been proposed [1, 2]. A fixed budget strategy for the distributed learning of the underlying nonparametric model has been introduced in [4].

So far [**?**, 1, 4], the GPR model for the Coherent-to-Diffuse power Ratio (CDR)-based distance and position estimation consisted of a mean function, which is identical to zero for all feature values and the squared exponential kernel. These choices are commonly made if no prior knowledge about the function to be estimated can be incorporated. In this contribution, we want to investigate the influence of alternative choices for the mean and the kernel function. Additionally, we introduce priors over the hyperparameters of these functions to express acquired prior knowledge, which leads to more robust identification of optimal hyperparameters. The influence of these choices on the performance of the algorithm is quantified in experiments within a simulated environment.

In the following, we denote matrices and vectors by boldface upper case and lower case letters, respectively. We mark training data by a tilde $\tilde{(\cdot)}$ and estimated quantities with a hat $\hat{(\cdot)}$. $[\cdot]_{i,j}$ denotes the ith row and jth column of a matrix and $[\cdot]_i$ denotes the ith entry of a vector.

In the remainder of the paper we introduce the used feature in Sec. 2 and we describe the learning-based distance estimation technique in Sec. 3. Experimental results are presented in Sec. 4 and the paper is concluded with Sec. 5.

2 DIFFUSENESS AS DISTANCE-RELATED FEATURE

We consider a single acoustic source emitting the signal $s(t)$, which is recorded by two microphones. The signal $x_i(t)$ observed at microphone $i \in \{1,2\}$ can be modeled as

$$x_i(t) = h_i(t) * s(t) + v_i(t) = \underbrace{h_{i,\text{early}}(t) * s(t)}_{c_i(t)} + \underbrace{h_{i,\text{late}}(t) * s(t) + v_i(t)}_{n_i(t)}, \tag{1}$$

where $*$ denotes linear convolution and $v_i(t)$ additive noise. The observed microphone signal can be separated into a coherent signal component $c_i(t)$ and a reverberant/noisy component $n_i(t)$, by dividing the Room Impulse Response (RIR) $h_i(t)$ into an early component $h_{i,\text{early}}(t)$ and a late component $h_{i,\text{late}}(t)$, which are convolved with the clean speech signal.

In [3], it has been shown that the ratio of the powers of these signal components, $c_i(t)$ and $n_i(t)$, yields a distance-dependent feature. To estimate this ratio, we use the CDR [15], which is based on an estimate of the complex-valued spatial coherence function of the microphone signals

$$\widehat{\Gamma}_{\text{x}}(l,\nu) = \frac{\hat{\Phi}_{x_1x_2}(l,\nu)}{\sqrt{\hat{\Phi}_{x_1x_1}(l,\nu)\hat{\Phi}_{x_2x_2}(l,\nu)}}, \tag{2}$$

where $l \in \{1,\dots,L\}$ denotes the time index and $\nu \in \{1,\dots,F\}$ the frequency index. The cross-Power Spectral Density (PSD) $\Phi_{x_1x_2}$ and the auto-PSDs $\Phi_{x_1x_1}$ and $\Phi_{x_2x_2}$ are estimated by recursive averaging over time

$$\hat{\Phi}_{x_ix_j}(l,\nu) = \lambda\hat{\Phi}_{x_ix_j}(l-1,\nu) + (1-\lambda)X_i(l,\nu)X_j^*(l,\nu) \quad \text{with} \quad i,j \in \{1,2\}, \tag{3}$$

where $\lambda \in [0,1]$ denotes a forgetting factor, $X_i(l,\nu)$ denotes the ith microphone signal in the Short-Time Fourier Transform (STFT) domain and $(\cdot)^*$ denotes complex conjugation. The coherence of a diffuse sound field can be modeled by [11]

$$\Gamma_{\text{n}}(\nu) = \frac{\sin(2\pi f_\nu d_{\text{mic}}/c)}{2\pi f_\nu d_{\text{mic}}/c}, \tag{4}$$

where f_ν denotes the physical frequency corresponding to frequency bin ν, d_{mic} the microphone spacing and c the speed of sound. Using the model for the diffuse sound field coherence (4) and the estimated microphone signal coherence (2), the following DOA-independent CDR estimator can be derived [15]

$$\begin{aligned}\widehat{\text{CDR}}(l,\nu) = \frac{1}{|\widehat{\Gamma}_{\text{x}}(l,\nu)|^2 - 1}\Big(&\Gamma_{\text{n}}(\nu)\,\text{Re}\{\hat{\Gamma}_{\text{x}}(l,\nu)\} - |\widehat{\Gamma}_{\text{x}}(l,\nu)|^2 \quad \dots \\ \dots \quad &- \sqrt{(\Gamma_{\text{n}}(\nu))^2\left(\text{Re}\{\widehat{\Gamma}_{\text{x}}(l,\nu)\}^2 - |\widehat{\Gamma}_{\text{x}}(l,\nu)|^2 + 1\right) - 2\Gamma_{\text{n}}(\nu)\,\text{Re}\{\widehat{\Gamma}_{\text{x}}(l,\nu)\} + |\widehat{\Gamma}_{\text{x}}(l,\nu)|^2}\Big).\end{aligned} \tag{5}$$

Finally, to obtain a robust feature, we average over the diffuseness $\frac{1}{\widehat{\text{CDR}}(l,\nu)+1}$ over all available time frames $1,\dots,L$ and the frequency interval indexed by $\nu_{\min},\dots,\nu_{\max}$

$$\zeta = \frac{1}{L(\nu_{\max}-\nu_{\min}+1)}\sum_{l=1}^{L}\sum_{\nu=\nu_{\min}}^{\nu_{\max}}\frac{1}{\widehat{\text{CDR}}(l,\nu)+1}. \tag{6}$$

Note that the values of the feature ζ are confined to the interval $[0,1]$ by construction.

3 LEARNING-BASED DISTANCE ESTIMATION

The relation of the feature ζ and the actual source-microphone distance r depends on several physical properties of the room which are unknown in practice. Therefore, we model the source-microphone distance r to be an unknown function of the observed feature, i.e., $g(\zeta)$ and we aim at learning this unknown relation $g : [0,1] \to \mathbb{R}_{\geq 0}$

for the room of interest by GPR, a nonparametric regression approach. The distances of the training set are modeled to be corrupted by additive Gaussian noise reflecting the uncertainty about the position of the source

$$r = g(\zeta) + \varepsilon \quad \text{with} \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{7}$$

Furthermore, for incorporating the prior knowledge about this relationship, the unknown function g is modeled to follow a Gaussian Process (GP)

$$g \sim \mathcal{GP}(m, k), \tag{8}$$

where $m : [0,1] \to \mathbb{R}_{\geq 0}$ denotes its prior mean function and $k : [0,1] \times [0,1] \to \mathbb{R}_{\geq 0}$ the chosen kernel function (see Sec. 3.2 and Sec. 3.3 for the discussion of different representatives). Both, m and k map from the feature domain $[0,1]$ to nonnegative numbers and depend on hyperparameters (see Sec. 3.1 for their optimal choice) which we stack in the vectors $\boldsymbol{\theta}_\mathrm{m}$ and $\boldsymbol{\theta}_\mathrm{k}$, respectively.

To infer an unseen distance r from an observed feature value ζ, we have to construct the posterior GP

$$\mathcal{GP}(m_\mathcal{D}, k_\mathcal{D}), \tag{9}$$

where $m_\mathcal{D}$ denotes the posterior mean function, $k_\mathcal{D}$ the posterior covariance function, and $\mathcal{D}$ the training set consisting of a vector of feature values $\tilde{\boldsymbol{\zeta}}$ and corresponding labels $\tilde{\mathbf{r}}$

$$\mathcal{D} = \left\{ \tilde{\boldsymbol{\zeta}}, \tilde{\mathbf{r}} \,\middle|\, \left[\tilde{\boldsymbol{\zeta}}\right]_i \in [0,1], [\tilde{\mathbf{r}}]_i \in \mathbb{R}_{\geq 0}, i \in \{1, \ldots, N\} \right\}. \tag{10}$$

We define the kernel matrix of the features of the training set as

$$\mathbf{K} = \left[k\left(\tilde{\zeta}_i, \tilde{\zeta}_j\right) \right]_{i,j} \quad \text{with} \quad 0 \leq i, j \leq N \tag{11}$$

and the kernel vector between the features of the training set and the new observation to be

$$\mathbf{k} = \mathbf{k}\left(\tilde{\boldsymbol{\zeta}}, \zeta\right) = \left[k\left(\tilde{\zeta}_i, \zeta\right) \right]_i \text{ with } 0 \leq i \leq N. \tag{12}$$

Any finite number of samples drawn from a GP follow a multivariate Gaussian Probability Density Function (PDF). Hence, we can express the joint density of the training labels $\tilde{\mathbf{r}}$ and the unseen distance r as

$$\begin{bmatrix} \tilde{\mathbf{r}} \\ r \end{bmatrix} \sim \mathcal{N}\left\{ \begin{bmatrix} \tilde{\boldsymbol{\mu}} \\ m(\zeta) \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{K}} + \sigma_\varepsilon^2 \mathbf{I} & \mathbf{k} \\ \mathbf{k}^\mathrm{T} & k(\zeta, \zeta) \end{bmatrix} \right\}, \tag{13}$$

where $[\tilde{\boldsymbol{\mu}}]_i = m([\tilde{\boldsymbol{\zeta}}]_i)$. By using the rules of Gaussian conditionals [13, p. 15-17], we obtain the estimate of the unseen distance $\hat{r}$ by evaluating the posterior mean function $m_\mathcal{D}$

$$\hat{r} = m_\mathcal{D}(\zeta) = m(\zeta) + \tilde{\mathbf{r}}^\mathrm{T} \left(\tilde{\mathbf{K}} + \sigma_\varepsilon^2 \mathbf{I}\right)^{-1} \mathbf{k}(\tilde{\boldsymbol{\zeta}}, \zeta). \tag{14}$$

Similarly, we obtain the posterior variance of the estimate

$$\hat{\sigma}_{\hat{r}}^2 = k_\mathcal{D}(\zeta, \zeta) = k(\zeta, \zeta) - \mathbf{k}^\mathrm{T}(\tilde{\boldsymbol{\zeta}}, \zeta) \left(\tilde{\mathbf{K}} + \sigma_\varepsilon^2 \mathbf{I}\right)^{-1} \mathbf{k}(\tilde{\boldsymbol{\zeta}}, \zeta). \tag{15}$$

3.1 Optimization of Hyperparameters

The joint posterior of the hyperparameters $\boldsymbol{\theta}_\mathrm{m}$ and $\boldsymbol{\theta}_\mathrm{k}$, given the training features $\tilde{\boldsymbol{\zeta}}$ and the training labels $\tilde{\mathbf{r}}$, can be written as

$$p\left(\boldsymbol{\theta}_\mathrm{m}, \boldsymbol{\theta}_\mathrm{k} \middle| \tilde{\mathbf{r}}, \tilde{\boldsymbol{\zeta}}\right) = \frac{p\left(\tilde{\mathbf{r}} \middle| \tilde{\boldsymbol{\zeta}}, \boldsymbol{\theta}_\mathrm{m}, \boldsymbol{\theta}_\mathrm{k}\right) p\left(\boldsymbol{\theta}_\mathrm{m}, \boldsymbol{\theta}_\mathrm{k} \middle| \tilde{\boldsymbol{\zeta}}\right)}{p\left(\tilde{\mathbf{r}} \middle| \tilde{\boldsymbol{\zeta}}\right)} \propto p\left(\tilde{\mathbf{r}} \middle| \tilde{\boldsymbol{\zeta}}, \boldsymbol{\theta}_\mathrm{m}, \boldsymbol{\theta}_\mathrm{k}\right) p(\boldsymbol{\theta}_\mathrm{m}, \boldsymbol{\theta}_\mathrm{k}), \tag{16}$$

where the marginal likelihood is defined by [13]

$$p(\tilde{\mathbf{r}}|\tilde{\boldsymbol{\zeta}},\boldsymbol{\theta}_\mathrm{m},\boldsymbol{\theta}_\mathrm{k}) = \mathcal{N}\left(\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}_\mathrm{m}),\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_\mathrm{k})\right). \tag{17}$$

Here, $\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}_\mathrm{m})$ and $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_\mathrm{k}) = \tilde{\mathbf{K}} + \sigma_\varepsilon^2\mathbf{I}$ emphasize that the prior mean vector and the covariance matrix depend on the hyperparameters $\boldsymbol{\theta}_\mathrm{m}$ and $\boldsymbol{\theta}_\mathrm{k}$, respectively. The dependency on $\tilde{\boldsymbol{\zeta}}$ has been omitted for clarity of presentation. We model the parameters of the mean function $\boldsymbol{\theta}_\mathrm{m}$ and the kernel function $\boldsymbol{\theta}_\mathrm{k}$ to be independent from each other

$$p(\boldsymbol{\theta}_\mathrm{m},\boldsymbol{\theta}_\mathrm{k}) = p(\boldsymbol{\theta}_\mathrm{m})p(\boldsymbol{\theta}_\mathrm{k}) = \prod_{i=1}^{n_\mathrm{m}} p\left([\boldsymbol{\theta}_\mathrm{m}]_i\right) \prod_{j=1}^{n_\mathrm{k}} p\left([\boldsymbol{\theta}_\mathrm{k}]_j\right), \tag{18}$$

where n_m and n_k denote the number of hyperparameters of the mean and kernel function, respectively. For identifying the optimum hyperparameters, we solve the following Maximum A Posteriori (MAP) optimization problem

$$\{\boldsymbol{\theta}_\mathrm{m}^\mathrm{opt},\boldsymbol{\theta}_\mathrm{k}^\mathrm{opt}\} = \underset{\boldsymbol{\theta}_\mathrm{m},\boldsymbol{\theta}_\mathrm{k}}{\operatorname{argmax}}\ p\left(\boldsymbol{\theta}_\mathrm{m},\boldsymbol{\theta}_\mathrm{k}|\tilde{\mathbf{r}},\tilde{\boldsymbol{\zeta}}\right). \tag{19}$$

Instead of solving (19) directly, it is common to maximize the log-posterior

$$\begin{aligned}\log p\left(\boldsymbol{\theta}_\mathrm{m},\boldsymbol{\theta}_\mathrm{k}|\tilde{\mathbf{r}},\tilde{\boldsymbol{\zeta}}\right) = &-\frac{1}{2}\left(\tilde{\mathbf{r}}-\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}_\mathrm{m})\right)^\mathrm{T}\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_\mathrm{k})^{-1}\left(\tilde{\mathbf{r}}-\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}_\mathrm{m})\right)\ldots \\ \ldots &-\frac{1}{2}\log\left(\det\left(\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_\mathrm{k})\right)\right) - \frac{N}{2}\log 2\pi + \sum_{i=1}^{n_\mathrm{m}}\log p\left([\boldsymbol{\theta}_\mathrm{m}]_i\right) + \sum_{j=1}^{n_\mathrm{k}}\log p\left([\boldsymbol{\theta}_\mathrm{k}]_j\right).\end{aligned} \tag{20}$$

To maximize the log-posterior w.r.t. the hyperparameters of the mean and the kernel function, a gradient ascent method is used. We start with the derivative of the prior mean function with respect to its hyperparameters $[\boldsymbol{\theta}_\mathrm{m}]_i$

$$\frac{\partial}{\partial[\boldsymbol{\theta}_\mathrm{m}]_i}\log p(\boldsymbol{\theta}_\mathrm{m},\boldsymbol{\theta}_\mathrm{k}|\tilde{\mathbf{r}},\tilde{\boldsymbol{\zeta}}) = \left(\tilde{\mathbf{r}}-\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}_\mathrm{m})\right)^\mathrm{T}\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_\mathrm{k})^{-1}\frac{\partial\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}_\mathrm{m})}{\partial[\boldsymbol{\theta}_\mathrm{m}]_i} + \frac{\partial\log p([\boldsymbol{\theta}_\mathrm{m}]_i)}{\partial[\boldsymbol{\theta}_\mathrm{m}]_i}. \tag{21}$$

The derivative of the log-posterior with respect to the kernel's hyperparameters $[\boldsymbol{\theta}_\mathrm{k}]_i$ results in

$$\frac{\partial}{\partial[\boldsymbol{\theta}_\mathrm{k}]_i}\log p(\boldsymbol{\theta}_\mathrm{m},\boldsymbol{\theta}_\mathrm{k}|\tilde{\mathbf{r}},\tilde{\boldsymbol{\zeta}}) = \frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{\beta}\boldsymbol{\beta}^\mathrm{T}-\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_\mathrm{k})^{-1}\right)\frac{\partial\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_\mathrm{k})}{\partial[\boldsymbol{\theta}_\mathrm{k}]_i}\right) + \frac{\partial\log p([\boldsymbol{\theta}_\mathrm{k}]_i)}{\partial[\boldsymbol{\theta}_\mathrm{k}]_i} \tag{22}$$

with $\boldsymbol{\beta} = \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_\mathrm{k})^{-1}\left(\tilde{\mathbf{r}}-\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}_\mathrm{m})\right)$.

In this contribution, we choose univariate Gaussian priors $\mathcal{N}(\theta|\theta_0,\sigma_\theta^2)$, with mean θ_0 and variance σ_θ^2, for the individual scalar parameters θ for simplicity, i.e., the prior PDFs in (21) and (22) become

$$p([\boldsymbol{\theta}_\mathrm{m}]_i) = \mathcal{N}([\boldsymbol{\theta}_\mathrm{m}]_i|[\boldsymbol{\theta}_{m,0}]_i,\sigma^2_{[\boldsymbol{\theta}_\mathrm{m}]_i}) \qquad \text{and} \qquad p\left([\boldsymbol{\theta}_\mathrm{k}]_j\right) = \mathcal{N}([\boldsymbol{\theta}_\mathrm{k}]_i|[\boldsymbol{\theta}_{k,0}]_i,\sigma^2_{[\boldsymbol{\theta}_\mathrm{k}]_i}). \tag{23}$$

The derivative of the log prior PDF w.r.t. the hyperparameter θ takes on the form

$$\frac{\partial}{\partial\theta}\log\mathcal{N}(\theta|\theta_0,\sigma_\theta^2) = \frac{\theta_0-\theta}{\sigma_{\theta_0}^2}, \tag{24}$$

where θ_0 is the expected parameter value and σ_θ^2 is a user-defined parameter. Note that the term (24) acts as a regularizer in the gradient ascent rules (21) and (22).

3.2 Choice of Mean Function

We discuss two different mean functions: the zero function and a linear function. The zero function

$$m(\zeta) = 0 \quad \forall \zeta \in [0,1] \tag{25}$$

is a common choice if no model of the function underlying the observed data is available and has been used in [1, 2, 4]. The discussion of the linear function

$$m(\zeta) = a_0 + a_1 \zeta \tag{26}$$

is motivated by the fact that the functional relation between the averaged diffuseness feature ζ and the corresponding distance r is monotonously increasing. Hereby, a_0 and a_1 denote the offset and the slope of the linear function, i.e., its hyperparameters.

3.3 Choice of Kernel Function

In this section, we investigate the applicability of a set of kernel functions k which all depend solely on the absolute difference between their two arguments $|\zeta_i - \zeta_j|$. All kernel functions are scaled by the signal variance β, a parameter which controls the deviation of the posterior mean function from the prior mean function. Also common to all kernel functions presented here is a positive length scale parameter α, which controls the width of the kernel and hence the smoothness of the posterior mean function.

We start our discussion with the squared exponential kernel function

$$k_{\mathrm{SE}}(\zeta_i, \zeta_j) = \beta \exp\left(-\frac{|\zeta_i - \zeta_j|^2}{2\alpha^2}\right) \quad \text{with} \quad \alpha > 0, \tag{27}$$

which is the most widely used kernel function. The Matérn class covariance functions

$$k_{\mathrm{Matérn}}(\zeta_i, \zeta_j) = \beta \frac{2^{1-\mu}}{\Gamma(\mu)} \left(\frac{\sqrt{2\mu}|\zeta_i - \zeta_j|}{\alpha}\right)^{\mu} K_{\mu}\left(\frac{\sqrt{2\mu}|\zeta_i - \zeta_j|}{\alpha}\right) \quad \text{with} \quad \mu, \alpha > 0 \tag{28}$$

are defined by the additional positive parameter μ. Hereby, Γ denotes the gamma function and K_{μ} the modified Bessel function of the second kind and μth order. Finally, the γ-exponential covariance function is defined, which contains the squared exponential covariance function as a special case

$$k_{\gamma}(\zeta_i, \zeta_j) = \beta \exp\left(-\left(\frac{|\zeta_i - \zeta_j|}{\alpha}\right)^{\gamma}\right) \quad \text{with} \quad 0 < \gamma \leq 2, \quad \alpha > 0, \tag{29}$$

with parameter γ.

4 EXPERIMENTAL STUDY

In the following section, we show results for experiments in a simulated enclosure and compare the performance of the algorithm using the probabilistic models outlined above.

4.1 Methodology

We used the RIR simulator [8] to generate RIRs of a single acoustic source observed by two microphones of spacing 0.2 m in two acoustic enclosures: Room 1 with dimensions $12\,\mathrm{m} \times 16\,\mathrm{m} \times 10\,\mathrm{m}$ and Room 2 with dimensions $6\,\mathrm{m} \times 5\,\mathrm{m} \times 3\,\mathrm{m}$. The N_{train} source positions of the training set are drawn randomly from an interval of source-microphone distances $[0\,\mathrm{m}, 3\,\mathrm{m}]$ with uniformly distributed angular spread of $\pm 10^{\circ}$ w.r.t. the microphone array normal. The $N_{\mathrm{eval}} = 200$ evaluation positions are generated analogously.

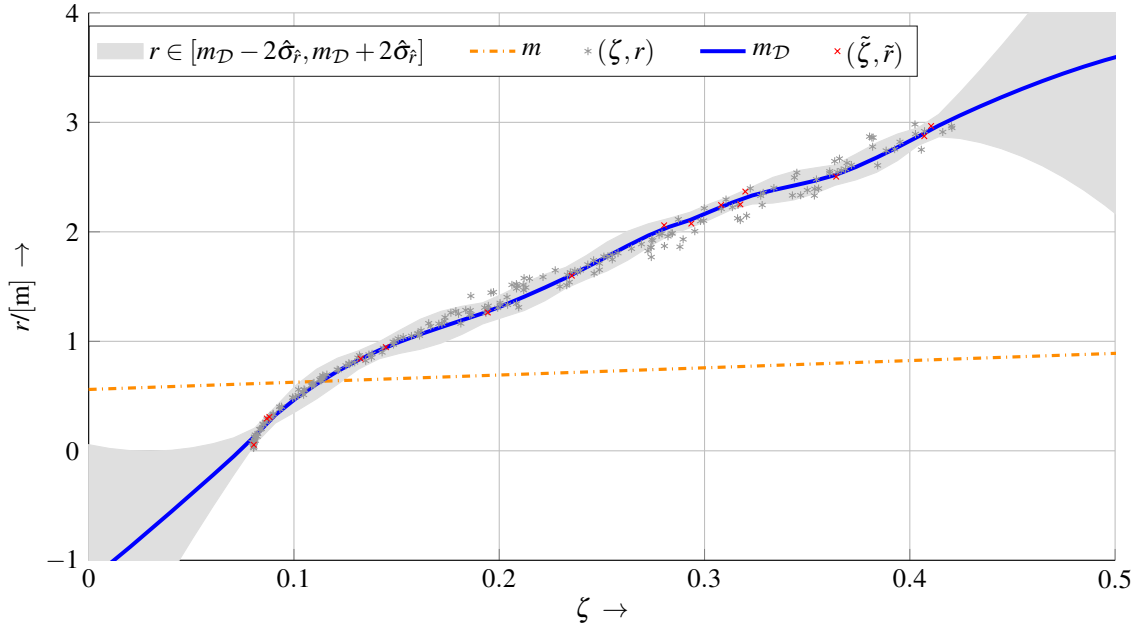

Figure 1. Example for a GP regression function (shown in blue) using the γ-exponential covariance function and a linear prior mean function (depicted as orange dashed line). Training (red crosses) and test data points (gray stars) have been generated in Room 1 under 30 dB SNR and $T_{60} = 0.8\,\mathrm{s}$.

To simulate microphone recordings, the RIRs are convolved with anechoic speech signals of 21.5 s duration and white Gaussian noise is added to simulate sensor noise. The microphone signals are transformed into the STFT domain using a von Hann window of length 25 ms and 10 ms shift at a sampling frequency of 16 kHz. The PSDs are estimated using $\lambda = 0.95$ and the diffuseness is calculated for the frequency interval $[125\,\mathrm{Hz}, 3468\,\mathrm{Hz}]$. All hyperparameters are initialized by 1 prior to optimization. We set priors on the hyperparameters of the covariance functions and set the corresponding parameters in (23) to $\beta_0 = 7$ and $\sigma^2_{\beta_0} = 1$, $\alpha_0 = 0.3$ and $\sigma^2_{\alpha_0} = 0.1$, and exclusively for the Matérn covariance function to $\mu_0 = 1.9$ and $\sigma^2_{\mu_0} = 0.6$.

We considered acoustic conditions of 10 dB, 20 dB, 30 dB SNR and $T_{60} \in \{0.4\,\mathrm{s}, 0.6\,\mathrm{s}, 0.8\,\mathrm{s}\}$ reverberation time. The number of training points has been chosen to be $N_{\text{train}} \in \{5, 10, 15, 20\}$. For each scenario, i.e., SNR, T_{60} and N_{train}, we repeated the simulations $N_{\text{runs}} = 10$ times and calculated the Absolute Error (AE) between the ground-truth distance r_{gt} and the estimated distance $\hat{r}$

$$e_{\text{AE}} = |\hat{r} - r_{\text{gt}}|. \tag{30}$$

4.2 Results

The aim of the proposed algorithm is the estimation of the nonparametric regression function $m_{\mathcal{D}} : [0,1] \rightarrow \mathbb{R}_{\geq 0}$, which enables the estimation of the acoustic source-microphone distance based on the observed feature value ζ. A typical result for the regression curve (shown as blue line) is depicted in Figure 1, which has been trained on samples (shown as red crosses) observed in Room 1 under 30 dB SNR and $T_{60} = 0.8\,\mathrm{s}$ reverberation time. The posterior standard deviation is depicted as a gray area around the posterior mean function, i.e., the regression function. It can be seen that the posterior standard deviation is high at positions where no training points are available and is low at positions where a lot of training points are observed. The prior mean function with optimized parameters is shown as orange dash-dotted line. The prior mean function is close to being a constant because all hyperparameters are optimized jointly and the data are already well-explained by the remaining modeling ability of the GP. Experiments showed that the choice of appropriate priors over the hyperparameters

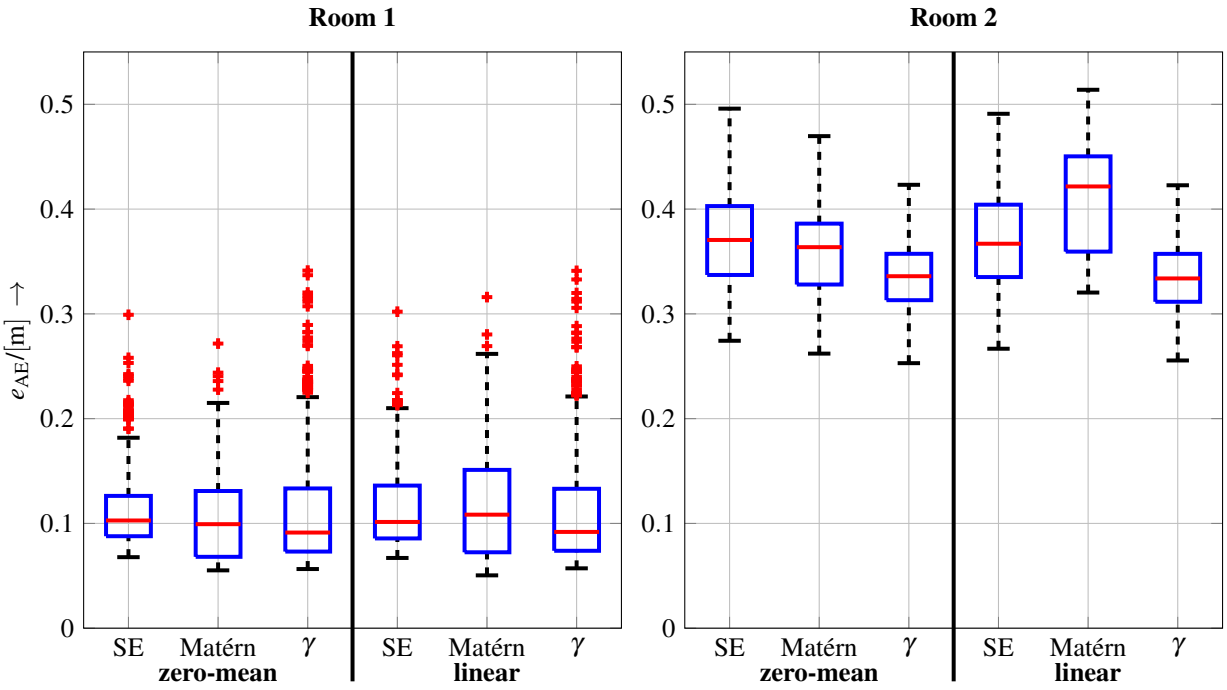

Figure 2. Boxplots of the AE in Room 1 on the left and in Room 2 on the right, for different acoustic parameters specified in the text. The results are separated w.r.t. the applied covariance functions (SE, Matérn, γ) and the applied prior mean function (zero-mean and linear).

yields more robust optimization results in the sense that a small deviation in the training data also yields only a small deviation in the regression function.

Results for the AE (30) e_{AE} in Room 1 and Room 2 are shown in Figure 2 for the applied covariance functions and the applied prior mean functions discussed in Sec. 3.2 and Sec. 3.3. First of all, it can be seen that the error for Room 1 is in general much smaller than for the corresponding results of Room 2. This can be explained by the fact that Room 1 is a very large room fulfilling the modeling assumptions of a diffuse noise field very well. Hence, the variance of the feature values corresponding to a certain distance is very small and the prediction of the trained regression function is very accurate. Due to dominant early reflections, the assumption of a diffuse sound field is not perfectly fulfilled for the smaller Room 2 yielding worse distance estimates.

In terms of the investigated covariance functions, the γ-exponential covariance function provides a small advantage over the competitors in Room 2 whereas the difference between the results is marginal for Room 1. The use of a linear prior mean function instead of a constant-zero prior mean function yields no further benefit. Hence, the approximative power of GP regression with a constant-zero prior mean function is sufficient for the modeling task. However, it should be noted that a non-zero prior mean function may be beneficial for regions where no training data exist, as the regression curve tends to return to the prior mean function if not supported by training data points. However, this effect is only relevant in special situations (unequal distribution of training and test data) and will not be discussed further here.

5 CONCLUSIONS

We investigated the influence of different probabilistic models, represented by the choice of covariance and prior mean functions and univariate Gaussian priors on their hyperparameters, on the acoustic source-microphone dis-

tance estimation performance of GP regression. To this end, we presented results based on a large number of experiments in different acoustic conditions and probabilistic models, obtained from microphone signals observed in a simulated environment.

It can be concluded that the choice of a nonzero prior mean function does not improve the results in general. Furthermore, the distance estimation results are only slightly affected by the choice of the covariance function, which was discernible for a room of regular size (Room 2) and not noticeable for a very large room (Room 1).

ACKNOWLEDGEMENTS

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

REFERENCES

[1] A. Brendel and W. Kellermann. Distance Estimation of Acoustic Sources using the Coherent-to-Diffuse Power Ratio Based on Distributed Training. In *IEEE Int. Workshop on Acoust. Signal Enhancement*, Tokyo, Japan, Sept. 2018.

[2] A. Brendel and W. Kellermann. Learning-based Acoustic Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio. In *European Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sept. 2018.

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

[4] A. Brendel and W. Kellermann. Distributed Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio. *IEEE J. of Select. Topics in Signal Process.*, pages 1–1, 2019.

[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] E. Georganti, T. May, S. van de Par, A. Harma, and J. Mourjopoulos. Speaker Distance Detection Using a Single Microphone. *IEEE Trans. on Audio, Speech, and Language Process.*, 19(7):1949–1961, Sept. 2011.

[7] E. Georganti, T. May, S. van de Par, and J. Mourjopoulos. Sound Source Distance Estimation in Rooms based on Statistical Properties of Binaural Signals. *IEEE Trans. on Audio, Speech, and Language Process.*, 21(8):1727–1741, Aug. 2013.

[8] E. A. P. Habets. Room Impulse Response Generator. Technical report, International Audio Laboratories, Erlangen, Germany, Sept. 2010.

[9] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda. Estimating Direct-to-Reverberant Energy Ratio Using D/R Spatial Correlation Matrix Model. *IEEE Trans. on Audio, Speech, and Language Process.*, 19(8):2374–2384, Nov. 2011.

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

[11] H. Kuttruff. *Room acoustics*. Spon Press/Taylor & Francis, London & New York, 5th edition, 2009.

[12] E. Larsen, C. Schmitz, C. Lansing, W. O'Brien, B. Wheeler, and A. Feng. Acoustic scene analysis using estimated impulse responses. In *Conf. Rec. of the Thirty-Seventh Asilomar Conf. on Signals, Syst. and Comput.*, pages 725–729, Pacific Grove, CA, USA, Nov. 2003.

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

[14] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. on Antennas and Propagation*, 34(3):276–280, Mar. 1986.

[15] A. Schwarz and W. Kellermann. Coherent-to-Diffuse Power Ratio Estimation for Dereverberation. *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, 23(6):1006–1018, June 2015.

[16] S. Vesa. Binaural Sound Source Distance Learning in Rooms. *IEEE Trans. on Audio, Speech, and Language Process.*, 17(8):1498–1507, Nov. 2009.