LEARNING-BASED ACOUSTIC SOURCE-MICROPHONE DISTANCE ESTIMATION USING THE COHERENT-TO-DIFFUSE POWER RATIO

Andreas Brendel and Walter Kellermann

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

ABSTRACT

We propose a method for estimating the distance between a sound source and a pair of recording microphones. The developed algorithm operates in the short-time Fourier transform domain and is based on estimates of the coherent-to-diffuse power ratio, which provides a measure for the amount of reverberation in each time-frequency bin. For a direct use of these estimates, precise knowledge on the room characteristics is necessary, which is in practice usually not available and hard to obtain. Therefore, we use a learning-based method, which adapts to the characteristics of the room in a training phase and estimates the source-microphone distance in a testing phase. The experiments comprise various setups with simulated and real data. It is shown that the proposed method generalizes well for different microphone positions and works robustly for different source signals, directions of arrival, reverberation times, and signal observation intervals. This leads to a high estimation accuracy at a low computational complexity with a small amount of training data.

Index Terms— Acoustic range estimation, Learning, Coherent-to-Diffuse Power Ratio

1. INTRODUCTION

The source-microphone distance is the length of the line-of-sight between an acoustic source and a recording microphone. Complementing the Direction of Arrival (DOA), it is one of the core features for estimating the position of an acoustic source [1] and knowing it is highly desirable for many audio signal processing applications involving distant talkers, e.g., teleconferencing [2], surveillance [3], hearing aids [4] and smart homes [5]. The estimation of the DOA has received much attention and many different methods have been established, e.g., based on correlation [6], blind system identification [7, 8] or clustering of phase differences [9]. On the other hand, the estimation of the source microphone distance is much less investigated. Two main classes of existing algorithms can be distinguished: Methods relying on prior information on the room properties [10, 11] and learning-based methods [13-17], which adapt to a specific room in a training phase. Prior knowledge about the room characteristics are, e.g., previously measured room impulse responses, as, e.g., used by [10]. Precise knowledge of the room characteristics, such as the absorption coefficients and the surface area of the walls, are used by [11]. However, such detailed prior knowledge is usually not available in practice and costly to obtain. Therefore, the second class of algorithms aims at learning these room characteristics before estimating the distances: A learning-based method for microphone array calibration was proposed by [12] and adapted by [13] for distance estimation. Unfortunately, this approach needs extensive training with white noise signals. A binaural method based on an estimate of the Direct-to-Reverberant Energy Ratio (DRR) was proposed by [14], which is capable of handling moving sound sources, but the localization performance degrades with increasing reverberation time T_{60} and the average localization error is reported to be up to $1\,\mathrm{m}$, which is too coarse for many applications. Monaural and binaural methods using several statistical measures of the recorded speech signals in combination with Gaussian classifiers and support vector machines were proposed in [15, 16]. A binaural learning-based distance estimation scheme using head-related transfer functions for robot audition was proposed by [17].

In this paper, we describe a learning-based approach for the estimation of the acoustic source-microphone distance. The proposed method is computationally efficient, robust against position changes of the recording microphones and accurate in terms of distance estimation error. Furthermore, the proposed method needs only a few training data points to fit the probabilistic model and produce reliable estimates in the trained acoustic environment in the testing phase. The resulting method is verified using simulated and measured Room Impulse Responses (RIRs) while varying acoustic conditions, amount of training data and source signals. Note that even though absolute distances are trained and evaluated in this contribution, relative distances between acoustic sources are sufficient for many practical applications, which simplifies the training drastically.

The proposed algorithm relies on the power ratio of the direct and the reflected sound components, which is an important cue for human listeners to estimate the distance of a sound source [18]. This ratio can be estimated using the spatial coherence [19], which yields a measure for the amount of reverberation in the observed signal, and can also be used for dereverberation [20]. Unlike [10] and [11], the proposed method does not rely on prior knowledge of the room parameters and applies a Gaussian classifier just as [12-14], but requires much less training data.

2. CDR ESTIMATOR

We consider one acoustic point source of unknown position, emitting an acoustic wideband signal in an enclosure. The acoustic signal is recorded by two microphones with spacing d_{mic} and is affected by reverberation, e.g., caused by reflections at the walls, and additive uncorrelated microphone noise. The m-th microphone signal $x_m(t)$ is modeled by a desired signal component $s_m(t)$ and an undesired signal component $n_m(t)$ representing noise and/or reverberation

$$x_m(t) = s_m(t) + n_m(t), \quad m \in \{1, 2\}. \tag{1}$$

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

$$\widehat{\mathrm{CDR}} = \frac{\Gamma_n \ \mathrm{Re}\left\{\hat{\Gamma}_x\right\} - \left|\hat{\Gamma}_x\right|^2 - \sqrt{\Gamma_n^2 \, \mathrm{Re}\left\{\hat{\Gamma}_x\right\}^2 - \Gamma_n^2 \left|\hat{\Gamma}_x\right|^2 + \Gamma_n^2 - 2\,\Gamma_n \ \mathrm{Re}\left\{\hat{\Gamma}_x\right\} + \left|\hat{\Gamma}_x\right|^2}}{\left|\hat{\Gamma}_x\right|^2 - 1} \tag{6}$$

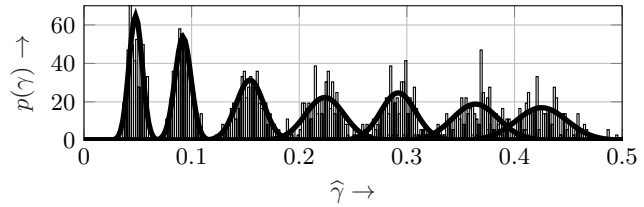

Fig. 1: Fitted Gaussian PDFs to averaged diffuseness values $\widehat{\gamma}$ for the distances $d \in \{0.2\mathrm{m}, 0.6\mathrm{m}, 1.0\mathrm{m}, \ldots, 2.6\mathrm{m}\}$ (from the left to the right).

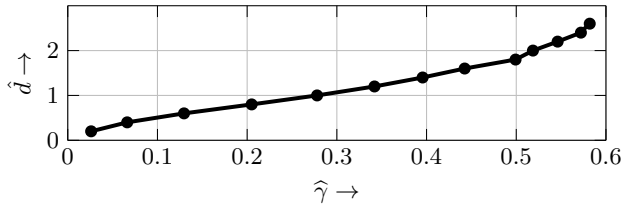

Fig. 2: Linear interpolation function fitted to averaged diffuseness values $\widehat{\gamma}$ for the distances $d_i \in \{0.2\mathrm{m}, 0.4\mathrm{m}, 0.6\mathrm{m}, \ldots, 2.6\mathrm{m}\}$.

The reverberant signal component is assumed to be generated by a diffuse sound field. In the Short-Time Fourier Transform (STFT) domain the microphone signals are denoted by $X_m(l,f)$ with l and f indexing time frame and frequency bin, respectively. The auto/cross Power Spectral Density (PSD) $\Phi_{x_m x_{m'}}(l,f)$ is estimated by recursive time averaging of the instantaneous microphone signal spectra

$$\hat{\Phi}_{x_m x_{m'}}(l,f) = \lambda \hat{\Phi}_{x_m x_{m'}}(l-1,f) + (1-\lambda) X_m(l,f) X^*_{m'}(l,f), \tag{2}$$

for $m, m' \in \{1,2\}$. The complex conjugate is denoted by $(\cdot)^*$ and λ is a smoothing factor. The complex spatial coherence function is given by

$$\hat{\Gamma}_x(l,f) = \frac{\hat{\Phi}_{x_1 x_2}(l,f)}{\sqrt{\hat{\Phi}_{x_1 x_1}(l,f)\hat{\Phi}_{x_2 x_2}(l,f)}}. \tag{3}$$

Based on this, the Coherent-to-Diffuse Power Ratio (CDR) between two omnidirectional microphones can be defined as [19]

$$\mathrm{CDR}(l,f) = \frac{\Gamma_n(l,f) - \Gamma_x(l,f)}{\Gamma_x(l,f) - \Gamma_s(l,f)}, \tag{4}$$

where $\Gamma_x(l,f)$, $\Gamma_s(l,f)$, $\Gamma_n(l,f)$ are the spatial coherence functions for the observations x, the desired signal s, and the reverberation/noise n, respectively. A high CDR corresponds to a low amount of reverberation in the respective time-frequency bin. Without loss of generality, the coherence function of the reverberant sound components is modeled here as a diffuse sound field

$$\Gamma_n(f) = \frac{\sin(2\pi f d_{\mathrm{mic}}/c)}{2\pi f d_{\mathrm{mic}}/c} \tag{5}$$

with speed of sound c. The estimator (6), which has been proposed in [21] (there called $\widehat{\mathrm{CDR}}_{\mathrm{prop3}}$), is used in this paper. Note that (6) yields real-valued estimates, whereas (4) generally does not. The time and frequency arguments l and f are omitted in (6) for brevity. Furthermore, $\mathrm{Re}\{\cdot\}$ denotes the real part and $|\cdot|$ the absolute value of a complex number, respectively, whereas $\hat{\Gamma}_x$ is the estimated spatial coherence of the two microphone signals in (3). The CDR estimator (6) is DOA-independent and is shown to be unbiased and robust [19], i.e., small deviations in the estimate of the spatial coherence $\hat{\Gamma}_x$ do not cause large deviations in $\widehat{\mathrm{CDR}}$. A detailed discussion of the properties of the estimator, as well as a comparison to different CDR estimators can be found in [19]. We use the diffuseness DIFF, defined in [19] as

$$\widehat{\mathrm{DIFF}}(l,f) = \frac{1}{\widehat{\mathrm{CDR}}(l,f) + 1}, \tag{7}$$

as the feature for the acoustic distance estimation in this contribution. To reduce the variance of the estimates and to obtain a single value as a feature for the estimation of the source-microphone distance, we average over all available time frames N_t and a frequency interval $[f_{\mathrm{min}}, f_{\mathrm{max}}]$ to obtain the averaged diffuseness

$$\widehat{\gamma} = \frac{1}{N_t(f_{\mathrm{max}} - f_{\mathrm{min}})} \sum_{l=1}^{N_t} \sum_{f=f_{\mathrm{min}}}^{f_{\mathrm{max}}} \widehat{\mathrm{DIFF}}(l,f). \tag{8}$$

Note that, when used with speech signals, $\widehat{\gamma}$ is in general dependent on the amount of speech pauses.

3. LEARNING-BASED SOURCE-MICROPHONE DISTANCE ESTIMATION

In this contribution, we use the averaged diffuseness $\widehat{\gamma}$ as a feature for the estimation of the distance between the center of the microphone pair and the source, as this is a feature which is easy to compute. $\widehat{\gamma}$ is not dependent on the source power, but increases with source-microphone distance as the power of the coherent signal components decreases with increasing source-microphone distance, whereas the diffuse signal power is constant [22]. Therefore, the estimated diffuseness constitutes a feature related to the source-microphone distance. However, the energy decay of the coherent signal components is also dependent on the room characteristics, e.g., the room volume or the reflection coefficients, which are unknown in practice and have to be estimated [23]. Hence, we propose to learn the mapping from the measured diffuseness values to a source-microphone distance for a given room via a classification algorithm based on training a Gaussian Mixture Model (GMM) and an algorithm based on linear interpolation.

3.1. Classification

We start by defining a set $\mathcal{D}$ of possible source microphone distances

$$\mathcal{D} = \{d_i \in \mathbb{R}_+ | d_i = d_0 + i\Delta d, i \in \mathbb{N}_0\}, \tag{9}$$

where Δd is the difference between two neighboring discrete distances of the set and d_0 is the minimum distance from the microphone pair. The algorithm is divided into a training phase and a testing phase. In the training phase, estimates $\widehat{\gamma}$ of the average diffuseness for known distances d_i of the grid $\mathcal{D}$ are collected and a GMM is trained for these labeled training data points. Similar learning methods have been frequently used in the literature [12-14]. In the testing phase, discrete source-microphone distances are inferred from classified $\widehat{\gamma}$ estimates in the same acoustic environment. The

set of $\widehat{\gamma}$ values obtained in the training phase is represented as a GMM, where each component corresponds to a discrete distance d_i and is defined by a mean μ_i and variance σ_i^2 describing the distribution of the estimated $\widehat{\gamma}$ for sources at this distance:

$$p_i\left(\widehat{\gamma}|d_i\right) = \mathcal{N}\left\{\widehat{\gamma}|\mu_i, \sigma_i^2\right\}. \tag{10}$$

We choose a uniform prior on d and obtain an approximation for the PDF of the averaged diffuseness on $\mathcal{D}$

$$p(\widehat{\gamma}) \approx \sum_{i=1}^{|\mathcal{D}|} p_i(\widehat{\gamma}|d_i), \tag{11}$$

where $|\mathcal{D}|$ is the cardinality of the set of possible distances. We denote the j-th data point of the averaged diffuseness of distance d_i with $\widehat{\gamma}_{i,j}$. With this definition the parameters of the Gaussian components can be estimated as follows [24]

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\widehat{\gamma}_{i,j} \quad \text{and} \quad \sigma_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(\widehat{\gamma}_{i,j} - \mu_i\right)^2, \tag{12}$$

where N_i is the number of measured training points $\widehat{\gamma}_{i,j}$ at distance d_i. An exemplary result is shown in Fig. 1 for a simulated room of dimensions $10\text{m} \times 8\text{m} \times 10\text{m}$ and a reverberation time of 1s [25]. The Gaussian components are trained on a grid with $d_0 = 0.2\text{m}$ and $\Delta d = 0.4\text{m}$. It can be seen, that the variances of the Gaussian components increase and simultaneously the difference between the means decrease for increasing $\widehat{\gamma}$. In the testing phase, the source microphone distances are estimated by choosing the Gaussian component, which yields the maximum likelihood for the estimated $\widehat{\gamma}$

$$\hat{d} = \underset{d_i \in \mathcal{D}}{\operatorname{argmax}}\; p_i\left(\widehat{\gamma}|d_i\right). \tag{13}$$

3.2. Interpolation

The proposed classification-based method produces accurate results if the length of the intervals between the distances used for training are small, i.e., the set of training distances is large. However, in a practical application, $|\mathcal{D}|$ will be very small and distances not coinciding with the learned grid points will cause large errors. To alleviate this problem, we introduce an interpolation approach based on linear interpolation in the following.

For describing the interpolation method, we first define the Lagrange interpolation polynomial of order one

$$L_i(\widehat{\gamma}) = \begin{cases} \frac{\widehat{\gamma}-\mu_{i-1}}{\mu_i-\mu_{i-1}} & \text{for } i > 1 \text{ and } \widehat{\gamma} \in [\mu_{i-1}, \mu_i) \\ \frac{\mu_{i+1}-\widehat{\gamma}}{\mu_{i-1}-\mu_i} & \text{for } i < |\mathcal{D}| \text{ and } \widehat{\gamma} \in [\mu_i, \mu_{i+1}] \\ 0 & \text{else} \end{cases}, \tag{14}$$

where $i = 1, \ldots, |\mathcal{D}|$ and μ_i is given by (12). Based on this, the interpolation estimate is given by evaluation of the interpolation function

$$\hat{d} = \sum_{i=1}^{|\mathcal{D}|} d_i L_i(\widehat{\gamma}). \tag{15}$$

An example for an interpolation function is shown in Fig. 2.

4. RESULTS

For all experiments we use two microphone recordings, obtained by convolving simulated or real RIRs with an anechoic speech signal. For the STFT, we choose a DFT length of 512, a von Hann window of length 25 ms and a frameshift of 10 ms. The smoothing factor for the estimation of the PSD is chosen to be $\lambda = 0.95$. This very strong averaging is chosen to reduce the influence of speech pauses and the nonstationarity of speech. The frequency interval $[f_{\text{min}}, f_{\text{max}}]$ in (8) is chosen to be $[125\,\text{Hz}, 3.5\,\text{kHz}]$ as this corresponds to the most significant part of the speech spectrum.

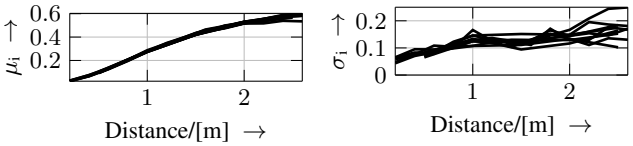

Fig. 3: **a)** Means μ_i and **b)** standard deviations σ_i of the GMM trained separately at all positions of Fig. 4 as a function of the source-microphone distance for grid resolutions $\Delta d = 0.2\text{m}$ and $\Delta d = 0.5\text{m}$.

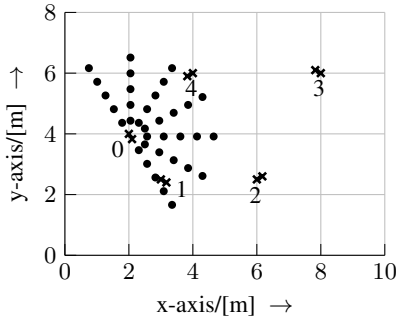

Fig. 4: Microphone positions for testing the generalization of the algorithm to different untrained positions (marked by crosses and numbers). The training positions for position 0 are marked by dots.

4.1. Simulated Data

First, we describe experimental results obtained with anechoic speech signals convolved with RIRs simulated by an image source model [26] and applying the room impulse generator [25]. We choose the room dimensions to be $10\,\text{m} \times 8\,\text{m} \times 10\,\text{m}$ and the reverberation time to be 1s. All sources and microphones are placed at a height of 2m. The microphone spacing is set to $d_{\text{mic}} = 0.2\,\text{m}$. The GMM is trained on a distance grid with $\Delta d = 0.2\,\text{m}/\Delta d = 0.5\,\text{m}$ and DOAs $\{0^\circ, 30^\circ, 60^\circ, \ldots, 180^\circ\}$ at each distance of the grid. In the testing phase, we generate source positions for testing the algorithm by drawing uniformly distributed distances $d \leq 2.6\,\text{m}$ and DOAs between 0° and 180°. To obtain a representative result, 100 random source positions (in general not coinciding with the distance grid $\mathcal{D}$), are evaluated for each training scenario. The median of the absolute distance estimation error $\tilde{e}$ is computed as a concise performance measure. Throughout the following experiments, larger errors are obtained for larger distances, which is caused by an increasing overlap of the Gaussian components (cf. Fig. 1) and decreasing distances between the estimated means of the GMM.

The estimation accuracy is tested for various source signals, namely two different noise (Gaussian/uniformly distributed) and five different speech signals and it is found that the results differed only marginally. Moreover, using identical or different signals for training and testing does not affect the estimation performance of the algorithm. Tab. 1 shows the median distance estimation error for the algorithm trained and evaluated at different positions of Fig. 4 respectively.

To investigate the generalization of the trained algorithm to different recording positions, the GMM trained at position 0 was eval-

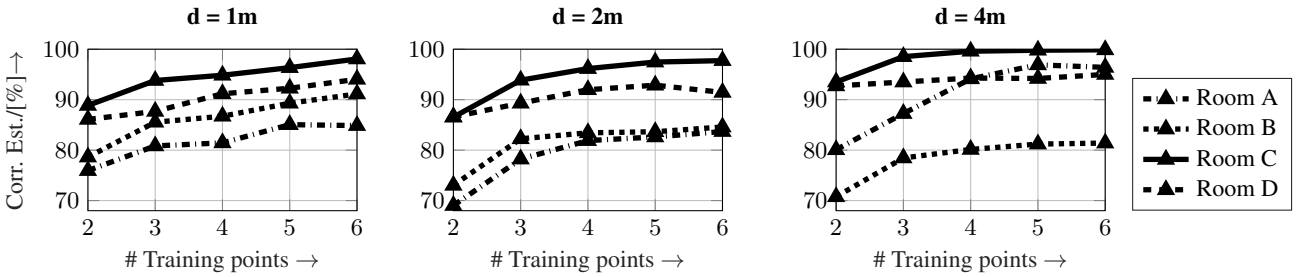

Fig. 5: Percentage of correctly identified source-microphone distances for real recordings in four rooms with different T_{60} values at distances of $\{1\,\mathrm{m}, 2\,\mathrm{m}, 4\,\mathrm{m}\}$ respectively, dependent on the number of training data points per Gaussian in the GMM.

		0	1	2	3	4
$\Delta d = 0.5\,\mathrm{m}$	Class.	16.0	14.3	14.2	15.5	15.2
	Interp.	6.3	6.6	6.4	8.1	6.6
$\Delta d = 0.2\,\mathrm{m}$	Class.	7.7	8.0	7.3	10.3	7.5
	Interp.	5.4	4.8	6.0	8.9	5.2

Table 1: Median absolute localization error $\tilde{e}$ in cm for different positions (see Fig. 4) for the classification algorithm as well as for the interpolation algorithm for different grid resolutions. The algorithm was trained and evaluated at the same positions respectively. Each experiment was repeated 100 times

		0	1	2	3	4
$\Delta d = 0.5\,\mathrm{m}$	Class.	16.0	16.7	14.3	15.8	15.5
	Interp.	6.3	7.0	6.4	9.6	6.4
$\Delta d = 0.2\,\mathrm{m}$	Class.	7.7	8.2	7.5	11.9	7.5
	Interp.	5.4	6.0	6.2	10.2	5.2

Table 2: Generalization of the algorithm trained at position 0 to different untrained microphone positions. $\tilde{e}$ in cm for the classification algorithm as well as for the interpolation algorithm for different grid resolutions. Each experiment was repeated 100 times.

uated at positions $\{1, 2, 3, 4\}$ (see Fig. 4). The results of this experiment are given in Tab. 2. The distance estimation errors are slightly higher in general compared to the results of Tab. 1. Furthermore, the GMM was trained at the positions $\{0, 1, 2, 3, 4\}$ and the resulting means μ_i and standard deviations σ_i of the GMM are plotted in Fig. 3 for grid resolutions $\Delta d = 0.2\,\mathrm{m}$ and $\Delta d = 0.5\,\mathrm{m}$. The means vary little between the different training positions for large source-microphone distances. Also the standard deviations follow the same trend for increasing distances. In summary, it can be concluded that the proposed method, once trained, generalizes well to different evaluation positions in the room. The performance of the algorithm is stable for different reverberation times T_{60}. The distance estimation error $\tilde{e}$ decreases for increasing T_{60} slightly, due to the fact that the diffuseness is increasing as the acoustic environment becomes more reverberant. Hence, the means of the Gaussian components have a wider spread over the interval of possible $\widehat{\gamma}$ values and are more distinct than for lower reverberation times. However, the performance differences between different T_{60} except for very low reverberation times are small. The effect of the duration of the microphone signals is small unless the signal duration is less than a second. The computational complexity for training and testing is dominated by the computation of (8) and thus essentially reduces to two FFTs and few simple algebraic operations, so that the overall computational complexity can be viewed as small.

4.2. Real Recordings

In this section, we evaluate our algorithm using recorded RIRs from two meeting rooms (A, B with $T_{60} = 0.2\,\mathrm{s}, 0.4\mathrm{s}$), respectively, a seminar room (Room C, $T_{60} = 0.7\,\mathrm{s}$) and a large foyer (Room D, $T_{60} = 3.5\,\mathrm{s}$). The data sets consist of 34, 34, 39 and 37 RIRs for rooms A, B, C and D respectively, measured from different source positions with source-microphone distances of $\{1\,\mathrm{m}, 2\,\mathrm{m}, 4\,\mathrm{m}\}$ and different DOAs. The distance between the microphones is chosen to be $d_{\mathrm{mic}} = 0.21\,\mathrm{m}$ in all recordings. The algorithm is trained on a subset of the available RIRs and is evaluated on the remaining ones. The number of correctly classified source distances is counted and divided by the number of all estimates for assessment of the performance of the algorithm. The relative amount of correctly classified distances is chosen as a measure due to the limited number of available RIRs. The training and testing phase is repeated 500 times by choosing the training subset randomly in each experiment. Fig. 5 shows the percentage of correctly classified source-microphone distances averaged over all experiments for different numbers of training points for the rooms described above.

The minimum training set consists in two samples. Even for this tiny training set, over 69% of the source positions are identified correctly. Note that for practical applications it is also possible to choose an arbitrary variance for the Gaussian components, such that one training data point would be enough. The amount of correctly identified source positions increases with the size of the training set. However, the increase is low for training sets larger than 4 samples, such that we can conclude that very few training points for each distance are sufficient to achieve a high estimation accuracy. Note that the experiments have shown that the proposed approach generalizes to different DOAs also in scenarios with measured RIRs.

5. CONCLUSION

We proposed a learning-based method for estimating the distance between an acoustic source and two recording microphones. The developed method has low computational complexity, needs only few training data points for each distance in a predefined grid, generalizes to different microphone positions and yields accurate distance estimates. The proposed approach is furthermore robust against changes in reverberation time, source signals, observed signal durations, DOAs and source-microphone distances. The effectiveness of the approach has been shown for real and simulated RIRs. Starting from a classification-based approach, an interpolation scheme was used to enhance the performance of the algorithm, especially for low grid resolution.

While the effectiveness of the algorithm is verified in this paper, a comparison with other algorithms that require much more training data is planned for future benchmarking.

6. REFERENCES

[1] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Process.*, vol. 107, pp. 54–67, Feb. 2015.

[2] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE ASSP Workshop on Appl. of Signal Process. to Audio and Acoustics*, New Paltz, NY, USA, Oct. 1997, pp. 1–4.

[3] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process. (ICASSP)*, Apr. 2009, pp. 165–168.

[4] M. Farmani, M.S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *IEEE Global Conf. on Signal and Inform. Process. (GlobalSIP)*, Dec. 2015, pp. 953–957.

[5] C. Lu, C. Wu, and L. Fu, "A Reciprocal and Extensible Architecture for Multiple-Target Tracking in a Smart Home," *IEEE Trans. on Syst., Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 1, pp. 120–129, Jan. 2011.

[6] J. 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.

[7] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The J. of the Acoust. Soc. of America*, vol. 107, no. 1, pp. 384, 2000.

[8] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "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.

[9] O. Schwartz, Y. Dorfan, E.A.P. Habets, and S. Gannot, "Multi-speaker DOA estimation in reverberation conditions using expectation-maximization," in *Proc. of the 15th Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Xi'an, China, Sep. 2016, pp. 1–5.

[10] E. Larsen, C.D. Schmitz, C.R. Lansing, W.D. O'Brien, B.C. Wheeler, and A.S. Feng, "Acoustic scene analysis using estimated impulse responses," in *Conf. Rec. of the Thirty-Seventh Asilomar Conf. on Signals, Syst. and Computers*, Pacific Grove, CA, USA, Nov. 2003, pp. 725–729.

[11] 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.*, vol. 19, no. 8, pp. 2374–2384, Nov. 2011.

[12] 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.

[13] 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.

[14] 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.

[15] 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.*, vol. 19, no. 7, pp. 1949–1961, Sep. 2011.

[16] 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.*, vol. 21, no. 8, pp. 1727–1741, Aug. 2013.

[17] F. Keyrouz, "Binaural range estimation using Head Related Transfer Functions," in *IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, San Diego, CA, USA, Sep. 2015, pp. 89–94.

[18] A.W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520, 1999.

[19] A. Schwarz and W. Kellermann, "Coherent-to-Diffuse Power Ratio Estimation for Dereverberation," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 23, pp. 1006–1018, Jun. 2015.

[20] J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone signalprocessing technique to remove room reverberation from speech signals," *The J. of the Acoust. Soc. of America*, vol. 62, no. 4, pp. 912–915, Oct. 1977.

[21] A. Schwarz and W. Kellermann, "Unbiased coherent-to-diffuse ratio estimation for dereverberation," in *Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Antibes-Juan les Pins, France, Sep. 2014, pp. 6–10.

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

[23] D. Markovic, K. Kowalczyk, F. Antonacci, C. Hofmann, A. Sarti, and W. Kellermann, "Estimation of Acoustic Reflection Coefficients Through Pseudospectrum Matching," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 22, no. 1, pp. 125–137, Jan. 2014.

[24] C. M. Bishop, *Pattern recognition and machine learning*, Information science and statistics. Springer, New York, 2006.

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

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