On Source-Microphone Distance Estimation Using Convolutional Recurrent Neural Networks

Tobias Gburrek, Joerg Schmalenstroeer, Reinhold Haeb-Umbach

Paderborn University, Department of Communications Engineering, Paderborn, Germany
Email: {gburrek, schmalen, haeb}@nt.uni-paderborn.de

Abstract

Several features computed from an audio signal have been shown to depend on the distance between the acoustic source and the receiver, but at the same time are heavily influenced by room characteristics and the microphone setup. While neural networks, if trained on signals representing a large variety of setups, have shown to deliver robust distance estimates from a coherent-to-diffuse power ratio (CDR) feature map at the input, we here push their modeling capabilities by additionally using the network as feature extractor. It is shown that distance estimation based on short-time Fourier transform (STFT) features can achieve a smaller estimation error and can operate on shorter signal segments compared to the previous CDR-based estimator.

1 Introduction

Localizing an acoustic source relative to the position and orientation of a compact microphone array amounts to estimating its direction of arrival (DoA) and its distance relative to the center of the array. Position estimation from acoustic signals finds multiple applications, such as steering a camera or realizing location-based services in the smart home. Furthermore, audio signal processing algorithms, e.g., acoustic beamforming [1] or blind source separation (BSS) based signal extraction [2, 3], can profit from position information. Additionally, tasks like synchronizing audio streams from distributed microphones [4, 5] or determining the geometry of wireless acoustic sensor networks (WASNs) [5, 6] require knowledge of the distances between sources and microphones in order to find a unique solution that reflects physical reality.

DoA estimation can be solved by employing knowledge about the array geometry and estimating the time-difference of arrival of the impinging audio signals at the microphones. There exist many methods for DoA estimation, e.g., those based on probabilistic models [7] or deep neural networks (DNNs) [8]. In contrast to that, acoustic distance estimation has found less attention. This may in part be because distance estimation from acoustic signals is heavily influenced by the room characteristics and therefore deemed less reliable. Thus, in unknown environments a distance estimator has either to infer the room characteristics from the microphone signals to adapt its estimation algorithm accordingly, or the distance estimation procedure has to be trained on a variety of rooms to generalize well to a target environment [9]. Hence, the existing distance estimation methods can be categorized into two classes.

First, there are methods which utilize prior knowledge about the room characteristics, e.g., previously measured room impulse responses (RIRs) [10] or the absorption coefficients of walls and other surfaces [11]. However, such prior knowledge is in general not available and furthermore, would be costly to obtain.

Second, there are learning-based methods, which typically use a training phase to adapt to a specific room. A common feature for these learning-based approaches is the direct-to-reverberant energy ratio (DRR) which was for example used to train a Gaussian mixture model (GMM) in [12], a Gaussian process (GP) in [13] or a DNN in [14]. For binaural setups other features in combination with classifiers have been proposed, e.g., Gaussian maximum-likelihood schemes based on the magnitude-squared coherence [15] or Gaussian classifiers and support vector machines trained on statistical measures, e.g., the standard deviation of interaural level differences of the recorded speech signals [16]. These learning-based approaches to distance estimation only work accurately if the acoustic environments of the training and deployment phase share very similar characteristics w.r.t. reverberation and room dimensions.

In this contribution we build upon the convolutional recurrent neural network (CRNN) based distance estimator we proposed in [9]. It uses a time-frequency representation of the coherent-to-diffuse power ratio (CDR) as input feature map and shows good generalization capabilities by exposing the CRNN to multiple different acoustic environments during training. Thus, there is no implicit need for an adaptation of the CRNN to the specific room in which it should be applied, if the room is similar to some of those seen during training.

The CDR is a feature extracted from the microphone signals' short-time Fourier transforms (STFTs) that does not reflect the full information present in the phase and magnitude of the STFTs. Prior works on CDR-based distance estimation [9, 17] have already shown that the distance estimator can benefit from additional features, e.g., DoA or acoustic environment information which both can be extracted from the STFT. So, presenting additional STFT features to a CRNN for distance estimation promises to outperform previous CDR-based approaches, since it gets direct access to all relevant information, some of which we highlight below.

The STFT coefficients well reflect the inter-channel level difference (ILD) information which is strongly linked to the source-microphone distance [16] and also the information contained in the CDR. Furthermore, the magnitude spectrum obtained from an STFT includes additional information which can be used to decide upon speech activity [18] as well as information about room characteristics [19]. Therefore, we propose here to use the STFT directly as a key input feature for a DNN.

In an extensive experimental study, we compare the STFT with the CDR and ILD features, including various combinations of all. Particular attention is paid to the dependence of the distance estimation accuracy on the length of the signal from which it is gleaned, since reducing the required signal length would improve the usefulness of distance estimation in dynamic scenarios.

The remainder of the paper is organized as follows: The investigated features for distance estimation are discussed in Sec. 2. Subsequently, our approach to CRNN-based distance estimation is introduced in Sec. 3. Finally, simulation results are presented in Sec. 4 before we end the paper by drawing conclusions in Sec. 5.

2 Investigated Features

We consider a closely spaced pair of microphones which is placed in a reverberant room. Moreover, we assume that this microphone pair records a single acoustic source, giving rise to the recorded microphone signals as follows:

$$y_i(t) = h_i(t) * x(t) + v_i(t) \text{ with } i \in \{1,2\}. \tag{1}$$

Here, $x(t)$ denotes the source signal, $h_i(t)$ the room impulse response modeling the sound propagation from the source position to the i-th microphone and $v_i(t)$ white sensor noise. The $*$ operator denotes a convolution. In the following we discuss features which represent distance-related quantities or side information being beneficial for distance estimation.

2.1 Coherent-to-Diffuse Power Ratio

Brendel et al. proposed to use the CDR for learning-based distance estimation [13], showing that the distance between microphone

and sound source is reflected by the power ratio between coherent and diffuse components of the recorded audio signals. Thereby, the coherent signal component is caused by the direct path and the early reflections, while the diffuse component comes from the late reflections and sensor noise.

Here, the DoA-independent CDR estimator which was proposed in [20, Eq. 12] is utilized to gather a time-frequency representation $\mathrm{CDR}(l,k)$, where l denotes the frame index and k the frequency bin index. Since bounding the feature to the interval $[0,1]$ is advantageous for the CRNN training, the CDR is transformed to the so-called diffuseness $\mathrm{D}(l,k)$ with

$$\mathrm{D}(l,k) = \frac{1}{1+\mathrm{CDR}(l,k)}, \tag{2}$$

which is finally used as input feature map of the CRNN.

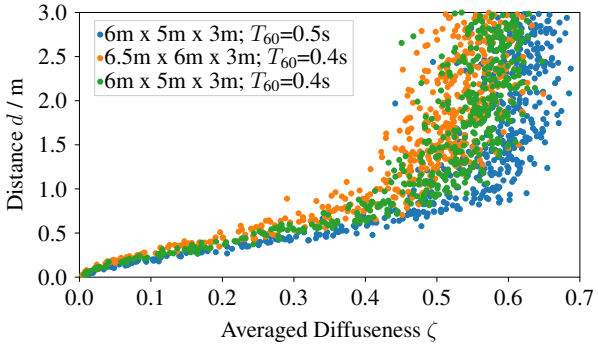

Figure 1: Relationship between the averaged diffuseness and the source-microphone distance: Each data point corresponds to a randomly drawn source-microphone constellation. The legend in the plot shows the dimensions of the considered rooms and the corresponding reverberation time T_{60}.

Due to the fact that the diffuseness is based on a power ratio it does not reflect all available information contained in the microphone signals. For example, information whether a coherent source is active in a time-frequency bin gets lost by calculating the power ratio. However, the CDR is only useful when a coherent source is active.

Furthermore, the relationship between the (averaged) diffuseness and the source-microphone distance d depends on the room characteristics, e.g., the reverberation time T_{60}, as shown in Fig. 1. Hereby, the averaged diffuseness ζ is defined by

$$\zeta = \frac{1}{T\cdot(k_{\max}-k_{\min}+1)} \sum_{l=0}^{T-1} \sum_{k=k_{\min}}^{k_{\max}} \mathrm{D}(l,k), \tag{3}$$

with T denoting the number of time frames and $k_{\min}$ and $k_{\max}$ corresponding to the limits of the considered frequency interval. Although it was shown in [9] that further side information, e.g., information about the room characteristics, is reflected by the time-frequency representation of the diffuseness to some extent, the estimator might benefit from a more direct representation of this information.

2.2 STFT-Related Features

The source-microphone distance can be estimated based on statistical information extracted from binaural signals as it was shown in [16]. Many of the statistical quantities for distance estimation proposed in literature can be derived directly from the STFT of the microphone signals. We briefly discuss two of them in the following before we explain how the STFT can be used directly as a feature.

2.2.1 Inter-Channel Level Differences

As described in [16], the standard deviation σ_{ILD} of the ILDs

$$\mathrm{ILD}(l,k) = \frac{|Y_1(l,k)|}{|Y_2(l,k)|} \tag{4}$$

calculated from the STFTs $Y_i(l,k)$, $i\in\{1,2\}$, of the microphone signals corresponds to a distance-related feature. Neglecting the sensor noise and expressing $\mathrm{ILD}(l,k)$ in dB results in

$$\begin{aligned}
\mathrm{ILD}^{(\mathrm{dB})}(l,k) &= 20\log_{10}\frac{|Y_1(l,k)|}{|Y_2(l,k)|} && (5)\\
&= 20\log_{10}\frac{|H_1(l,k)X(l,k)|}{|H_2(l,k)X(l,k)|} && (6)\\
&= 20\log_{10}|H_1(l,k)| - 20\log_{10}|H_2(l,k)| && (7)\\
&= H_1^{(\mathrm{dB})}(l,k) - H_2^{(\mathrm{dB})}(l,k), && (8)
\end{aligned}$$

with $X(l,k)$ denoting the STFT of the source signal and $H_i(l,k)$, $i\in\{1,2\}$, the STFTs of the RIRs.

Based on this representation of $\mathrm{ILD}^{(\mathrm{dB})}(l,k)$, it is shown in [16] that the standard deviation of $\mathrm{ILD}^{(\mathrm{dB})}(l,k)$ is a function of the source-microphone distance d:

$$\begin{aligned}
\sigma_{\mathrm{ILD}} &= \sqrt{\frac{\sum_{l=0}^{T-1}\sum_{k=k_{\min}}^{k_{\max}}\left(\mathrm{ILD}^{(\mathrm{dB})}(l,k)-\mu_{\mathrm{ILD}^{(\mathrm{dB})}}\right)^2}{T\cdot(k_{\max}-k_{\min}+1)}} && (9)\\
&= f\left(\sigma_1^2(d)+\sigma_2^2(d)\right), && (10)
\end{aligned}$$

whereby $\mu_{\mathrm{ILD}^{(\mathrm{dB})}}$ corresponds to the mean of $\mathrm{ILD}^{(\mathrm{dB})}(l,k)$. $\sigma_i^2(d)$, $i\in\{1,2\}$, denotes the variance of the RIR corresponding to the i-th microphone which is defined as

$$\sigma_i^2(d) = \frac{\sum_{l=0}^{T-1}\sum_{k=k_{\min}}^{k_{\max}}\left(H_i^{(\mathrm{dB})}(l,k)-\mu_{H_i^{(\mathrm{dB})}}\right)^2}{T\cdot(k_{\max}-k_{\min}+1)}, \tag{11}$$

with $\mu_{H_i^{(\mathrm{dB})}}$ denoting the mean of $H_i^{(\mathrm{dB})}(l,k)$.

In [21] the relationship between σ_i^2 and the distance d was derived which is given by

$$\sigma_i^2(d) = \frac{1+2r}{(1+r)^2}, \text{ with } r=\frac{d_c^2}{d^2}. \tag{12}$$

Here, d_c corresponds to the critical distance of the room being dependent on the room volume and the reverberation time T_{60}. As shown in Fig. 2 the relation between σ_{ILD} and the source-microphone distance exhibits a similar behavior as the relation between the averaged diffuseness and the source-microphone distance however being less dependent on the room characteristics.

We provide the raw ILDs expressed in dB as a single feature map to the CRNN and let the CRNN extract the distance-related high-level features. Due to the fact that the source activity information which is contained in the magnitude of the STFT [18] gets lost when the ILD features are calculated the distance estimator might profit from the magnitude of the STFT of one microphone signal as an additional feature map.

2.2.2 Inter-Channel Phase Differences

Inter-channel phase difference (IPD) features, i.e., the difference of the phases of the microphone signals' STFTs, provide useful side information, e.g., information about the reverberation time T_{60} [22]. This knowledge of room characteristics is helpful because most distance-related features significantly depend on room characteristics. In addition to that, the phase differences deliver DoA information which can be beneficial for distance estimation [17]. Before providing them to the CRNN the sine and cosine of the IPDs are taken as proposed in [23]. This results in two feature maps emphasizing those frequency bands which best show the phase differences compared to raw IPD features.

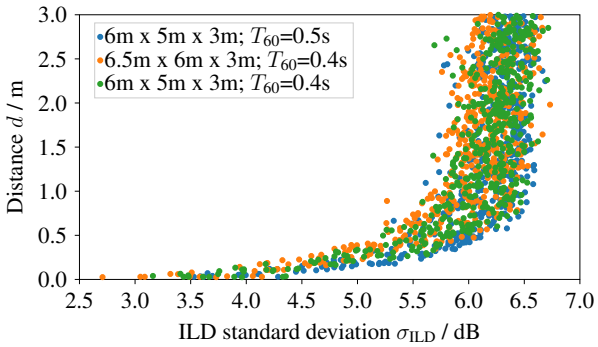

Figure 2: Relationship between the standard deviation of the ILDs and the source-microphone distance: Each data point corresponds to a randomly drawn source-microphone constellation. The legend in the plot shows the dimensions of the considered rooms and the corresponding reverberation time T_{60}.

2.2.3 Direct Usage of the STFT

The STFT of the microphone signals can be interpreted as a very general representation of distance information and useful side information. While the use of the STFT ensures that no information is lost, its dependence on the distance is not as apparent as it is with hand-crafted features such as the CDR. Therefore, the first part of the CRNN proposed in the next section is designed to act as a powerful feature extractor to gather the information important for distance estimation from the STFT coefficients.

For this purpose, the STFTs of both microphone signals are presented in terms of their magnitudes and phases to the CRNN. Using the STFTs of the microphone signals as features results in four input feature maps. From the magnitude maps information about source activity and distance-related ILDs can be inferred, and the phase maps contain beneficial side information about the DoA and the reverberation time T_{60}.

3 CRNN-Based Distance Estimation

The CRNN-based distance estimator used in this paper is an adapted version of the distance estimator we proposed in [9]. Both handle distance estimation as a classification task. Therefore, the distances are quantized with a granularity of 0.1 m and the working range is restricted to distances below a maximum distance $d_{\max}$. To learn a mapping between input features and distances that also generalizes to unseen acoustic environments, the CRNN is exposed to data from different acoustic environments during training. The architecture of the chosen CRNN is summarized in Tab. 3.

The CRNN gets a $B \times M \times F \times T$ dimensional time-frequency representation of the microphone signals as input, which corresponds to the results of the feature extraction process explained in Sec. 2. Here, B denotes the size of the mini-batches, M the number of input feature maps, F the number of frequency bins and T the number of time frames. For example, M=4 holds if the STFT (magnitude and phase of the two microphone signals) is directly used as input feature.

First, there is a convolutional module consisting of a 2D convolutional neural network (CNN) intended for extracting high-level feature maps, followed by a 1D CNN intended to combine the information of neighboring frames by processing all of their frequencies. Each of the three blocks of the 2D CNN comprises two 2D convolutional layers with 64 channels using a kernel of size 7×3 and a stride of one. After the two 2D convolutional layers a max pooling layer with stride four along the frequency dimension follows. Unlike the CRNN architecture that we proposed in [9], no pooling is performed along the time dimension to allow shorter signal segments as input to the estimator. The 1D CNN consists of two 1D convolutional layers with 512 channels, a kernel of size 3 and a stride of one.

Block	Output shape
Feature Extraction	$B \times M \times F \times T$
$2 \times$ Conv2D$(7 \times 3; 64)$ MaxPool2D(4×1)	$B \times 64 \times F \times T$ $B \times 64 \times \lfloor F/4 \rfloor \times T$
$2 \times$ Conv2D$(7 \times 3; 64)$ MaxPool2D(4×1)	$B \times 64 \times \lfloor F/4 \rfloor \times T$ $B \times 64 \times \lfloor F/16 \rfloor \times T$
$2 \times$ Conv2D$(7 \times 3; 64)$ MaxPool2D(4×1) Reshape	$B \times 64 \times \lfloor F/16 \rfloor \times T$ $B \times 64 \times \lceil F/64 \rceil \times T$ $B \times 64 \cdot \lceil F/64 \rceil \times T$
$2 \times$ Conv1D$(3; 512)$	$B \times 512 \times T$
$2 \times$ GRU(256)	$B \times 256$
$\text{fc}_{\text{ReLU}}(256)$ $\text{fc}_{\text{Softmax}}(C)$	$B \times 256$ $B \times C$

Table 1: Architecture of the proposed CRNN: Each conv{1,2}D layer includes batch normalization and ReLU activation. Dropout with a probability of 0.5 is applied to the outputs of the hidden layers of the recurrent and the fully connected part. $\lfloor \cdot \rfloor$ corresponds to the flooring operator.

The sequence of feature vectors generated by the convolutional module is processed by a recurrent module consisting of two gated recurrent unit (GRU) layers with 256 units per layer. Thereby, the recurrent module is intended to extract temporal information from the feature vector sequence and to summarize all information of the processed signal segment for a final single decision on the distance for the entire segment. Therefore, only the last output vector is forwarded to the final classification layers.

The final fully connected module contains a hidden layer with 256 units and ReLU activation function and a final classification layer with C units and Softmax activation function. Here, C corresponds to the number of considered distance classes.

4 Experiments

We use a simulated data set corresponding to the setup in Fig. 3 to evaluate the performance of the distance estimators utilizing the different features which are described in Sec. 2. The data set is split into a training set consisting of 100k, a validation set of 1000 and an evaluation set of 10k source microphone pair constellations. Each of the source microphone pair constellations is placed in a room whose width and depth are randomly drawn from $[5\,\text{m}, 7\,\text{m}]$. All rooms have a fixed height of 2.4 m and all microphones and acoustic sources are placed on the same height of 1.15 m. The reverberation time T_{60} of the rooms is uniformly drawn at random from $[0.2\,\text{s}, 0.7\,\text{s}]$.

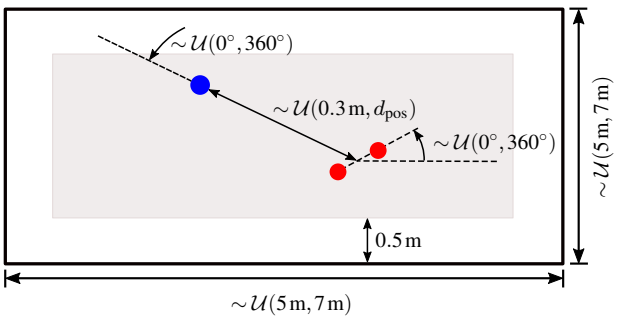

Figure 3: Simulated setup: The gray area visualizes the area in which the microphones (red dots) and the acoustic source (blue dot) are randomly placed.

To achieve an approximately uniform distribution of the distances across all random constellations of the data set, the following procedure is used: First, the two microphones having a

spacing of 5 cm are placed with a random position and orientation within the considered area and the distance is uniformly drawn at random from $[0.3\,\mathrm{m}, d_{\mathrm{pos}}]$. Here, d_{pos} corresponds to the minimum of d_{max}=5 m and the largest possible distance so that the source could be placed in the considered area. Subsequently, the DoA is drawn at random. If the acoustic source would have to be placed outside the considered area for the drawn distance and DoA, the DoA is increased until the source position is within the considered area. Note that the desired uniform distribution of the distances is achieved via a modification of the DoA distribution, to which we pay less importance. For each source microphone pair constellation the RIRs are generated by the image source method utilizing the implementation of [24].

The acoustic source signals are random samples from the TIMIT database [25] that are trimmed to the specific length under consideration for the experiments. Samples for training purposes are taken from the TIMIT train set, while test samples are from the TIMIT test set. At test time 100 speech samples are randomly drawn and reverberated for each source microphone pair constellation of the evaluation set. Subsequently, each reverberated speech sample is cut into segments of the considered length, whereby all resulting segments are used for the experimental evaluation.

The distance is linearly quantized with a granularity of 0.1 m resulting in C=48 distance classes. All distance estimators are trained for 500k iterations using the Adam optimizer [26] with a mini-batch size of B=32 and a learning rate of $3 \cdot 10^{-4}$. After training the best performing checkpoint w.r.t. the mean absolute error (MAE) of the distance estimates on an independent validation set is chosen. The STFT used for the extraction of all investigated features utilizes a Blackman window with a length of 25 ms and a shift of 10 ms. As described in [13] a forgetting factor of λ=0.95 is used to estimate the power spectral densities which are needed to calculate $\mathrm{CDR}(l, k)$.

We employ the MAE and the accuracy of the distance classification as performance metrics. Here, the MAE, i.e., the average absolute difference between the estimated distance and the ground truth distance before quantization, is calculated over all reverberated speech segments. Although the fine-granular quantization of distances can quickly result in a confusion with a nearby distance class, such a confusion still results in a small distance error. Therefore, in addition to the values for the classical definition of accuracy, we also state values for an extended accuracy that allows a confusion with the next-closest distance class w.r.t. the ground truth distance before quantization.

Feature	**Seg. length** / ms		**MAE** / m	**Acc.** / %	
	train	**test**		normal	extended
CDR	256	256	0.23	68.1	77.3
CDR	256	512	0.13	79.0	87.9
CDR	256	1024	0.09	83.2	91.9
CDR	1024	256	0.48	57.8	66.5
CDR	1024	512	0.15	77.9	86.8
CDR	1024	1024	0.07	86.4	94.6
STFT	256	256	0.15	76.6	85.3
STFT	256	512	0.07	86.4	94.2
STFT	256	1024	0.05	89.8	97.0
STFT	1024	256	0.36	65.7	73.8
STFT	1024	512	0.10	84.1	91.6
STFT	1024	1024	0.04	91.1	97.5

Table 2: Influence of the signal segment length on the performance of the distance estimator.

Tab. 2 shows the effect of the signal segment length on distance estimation for the selected metrics. As expected a reduction of the segment length leads to worse distance estimates. Furthermore, the effect intensifies if the distance estimator is trained on longer segments. The first observation can be explained by the fact that the proportion of time-frequency bins without speech activity is larger for shorter segments, while the latter observation is dedicated to a mismatch of this proportion between training and test data. We can conclude that longer segments are beneficial for estimation accuracy, but that it does not make sense to employ longer segments in training than are used in test.

A direct comparison between CDR and STFT features shows that under the same training and test conditions, the STFT features result in an improved performance in all considered cases. For example, an STFT-based estimator (trained on 256 ms long segments) can achieve an equally good performance on 512 ms long segments as a CDR-based estimator (trained on 1024 ms long segments) on 1024 ms long segments. To get a more detailed insight into the reasons for the differences, experiments with additional features, i.e., ILD features, IPD features, magnitude and phase, are considered in the following.

Feature	**MAE** / m	**Acc.** / %	
		normal	extended
CDR	0.23	68.1	77.3
ILD	0.32	61.5	69.9
STFT	0.15	76.6	85.3
ILD + Mag.	0.25	65.3	75.5
ILD + Mag. + Phase	0.16	75.0	84.0
ILD + Mag. + IPD	0.17	73.9	82.7
CDR + Mag.	0.18	72.4	82.3
CDR + Mag. + Phase	0.15	77.5	86.3
CDR + Mag. + IPD	0.15	76.5	85.4
CDR + STFT	0.13	78.8	87.2

Table 3: Comparison of the investigated input features for a signal segment length of 256 ms. 'Mag.' denotes that the magnitude of the STFT of one channel is added. 'Phase' denotes that the phases of the STFTs of both microphone signals are added.

From Tab. 3 it can be seen that the STFT features carrying additional information beside distance information outperform the CDR and ILD features. By successively adding additional information, e.g., source activity information reflected by the magnitude of the STFT, the MAE of the CDR- and ILD-based estimators can be decreased. Thereby, the phase information of the microphone signal's STFTs, which is among other things an indicator of the reverberation time T_{60}, gains the largest improvement of the distance estimation performance. The best performance results from the combination of the CDR feature and the STFT features, which provides complementary distance information as well as useful side information.

5 Conclusions

In this paper, we presented an experimental study of CRNN-based source-microphone distance estimation using a neural network for feature extraction in addition to precomputed high-level features. From the STFT representation of the recorded audio signals, the CRNN manages to extract both distance-related information and additional side information, e.g., source activity and acoustic environment information, useful for the task. The presented STFT-based distance estimator is able to reduce the average distance error, compared to a recently proposed CDR-based distance estimator as well as ILD-based methods, by almost and by more than a factor of two, respectively. Moreover, the newly proposed distance estimator can work with shorter speech segment lengths, which improves its overall usefulness in dynamic scenarios.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project 282835863. We would like to thank our student Nils Horsmann for conducting first experiments on the usage of the STFT as input feature for distance estimation.

References

[1] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[2] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Combined lcmv-trinicon beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 320–332, 2017.

[3] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3545–3558, 2020.

[4] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), Oct. 2009.

[5] S. Wozniak and K. Kowalczyk, "Passive Joint Localization and Synchronization of Distributed Microphone Arrays," *IEEE Signal Processing Letters*, vol. 26, pp. 292–296, Feb 2019.

[6] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "Geometry calibration in wireless acoustic sensor networks utilizing doa and distance information," *Accepted for publication in EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[7] L. Drude, F. Jacob, and R. Haeb-Umbach, "DOA-estimation based on a complex Watson kernel method," in *Proc. European Signal Processing Conference (EUSIPCO)*, (Nice, France), Aug 2015.

[8] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 136–140, 2017.

[9] T. Gburrek, J. Schmalenstroeer, A. Brendel, W. Kellermann, and R. Haeb-Umbach, "Deep Neural Network based Distance Estimation for Geometry Calibration in Acoustic Sensor Networks," in *Proc. European Signal Processing Conference (EUSIPCO)*, (Amsterdam, The Netherlands), jan. 2020/2021.

[10] E. Larsen, C. Schmitz, C. Lansing, W. O'Brien, B. Wheeler, and A. Feng, "Acoustic scene analysis using estimated impulse responses," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 1, pp. 725–729 Vol.1, 2003.

[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 Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, 2011.

[12] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.

[13] A. Brendel and W. Kellermann, "Learning-based acoustic source-microphone distance estimation using the coherent-to-diffuse power ratio," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2018.

[14] M. Zohourian, J. Stinner, and R. Martin, "Speaker distance estimation using binaural hearing aids and deep neural networks," in *23rd International Congress on Acoustics*, 2019.

[15] S. Vesa, "Binaural sound source distance learning in rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1498–1507, 2009.

[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 Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1727–1741, 2013.

[17] A. Brendel, I. Altmann, and W. Kellermann, "Acoustic source position estimation based on multi-feature gaussian processes," in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2019.

[18] J. Heitkaemper, J. Schmalenstroeer, and R. Haeb-Umbach, "Statistical and neural network based speech activity detection in non-stationary acoustic environments," in *INTERSPEECH 2020 Virtual Shanghai China*, 2020.

[19] Y. Khokhlov, A. Zatvornitskiy, I. Medennikov, I. Sorokin, T. Prisyach, A. Romanenko, A. Mitrofanov, V. Bataev, A. Andrusenko, M. Korenevskaya, and O. Petrov, "R-Vectors: New Technique for Adaptation to Room Acoustics," in *Proc. Interspeech 2019*, pp. 1243–1247, 2019.

[20] A. Schwarz and W. Kellermann, "Unbiased coherent-to-diffuse ratio estimation for dereverberation," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, (Antibes, French Riviera), sep. 2014.

[21] K. J. Ebeling, "Influence of direct sound on the fluctuations of the room spectral response," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1206–1207, 1980.

[22] M. Lee and J. Chang, "Deep neural network based blind estimation of reverberation time based on multi-channel microphones," *Acta Acustica united with Acustica*, vol. 104, pp. 486–495, May 2018.

[23] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2018.

[24] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.

[26] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. International Conference on Learning Representations (ICLR)*, (Banff, Canada), apr. 2014.