CHALMERS
UNIVERSITY OF TECHNOLOGY

Blind Estimation of Spatial Room Impulse Responses Using a Pseudo Reference Signal

Citation for the original published paper (version of record):
Deppisch, T., Ahrens, J., Garí, S. et al (2024). Blind Estimation of Spatial Room Impulse Responses Using a Pseudo Reference Signal. 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops, ICASSPW 2024 - Proceedings: 470-474.
http://dx.doi.org/10.1109/ICASSPW62465.2024.10626717

BLIND ESTIMATION OF SPATIAL ROOM IMPULSE RESPONSES USING A PSEUDO REFERENCE SIGNAL

Thomas Deppisch, Jens Ahrens[*]

Chalmers University of Technology,
412 96 Gothenburg, Sweden
{thomas.deppisch, jens.ahrens}@chalmers.se

Sebastià V. Amengual Garí, Paul Calamia

Reality Labs Research, Meta,
Redmond, WA 98052, USA
{samengual, pcalamia}@meta.com

ABSTRACT

In auditory augmented reality applications, virtual sound sources can be added to a real-world acoustic environment by processing each source signal with a spatial room impulse response (SRIR) to render acoustic characteristics of the environment, and with a set of head-related transfer functions to create binaural headphone signals. The SRIR of a user's environment is typically unknown and needs to be estimated. We propose a method to estimate the SRIR blindly from speech signals captured with a microphone array. The blind estimation task is transformed into a non-blind one using a pseudo reference signal that is obtained from the array signals via dereverberation and beamforming. The SRIR is then estimated using a frequency-domain multichannel Wiener filter with the pseudo reference as the input and the array signals as the desired signals. In contrast to conventional methods, the proposed method is able to successfully estimate SRIRs of realistic lengths at a sampling rate that supports the entire audible frequency range. Results from 200 simulated and 16 measured SRIRs show that the estimates from the proposed method reproduce the reverberation time and the direct-to-reverberant energy ratio with low error, outperforming a baseline method that does not use dereverberation.

***Index Terms*—** Augmented Reality, Blind System Identification, Dereverberation, Microphone Array, Room Impulse Response

1. INTRODUCTION

In auditory augmented reality applications, virtual sound objects may be added to a real-world environment that the user can freely explore, typically while wearing non-occluding headphones to limit the distortion of real-world sounds. A specific use case is augmented telepresence which aims at making remote speech appear in the local acoustic environment and vice versa. A virtual sound source can be created by processing a source signal with the binaural transfer function from a position in the real-world environment to the listener's ears. The binaural transfer function comprises the acoustic room transfer function, or its time-domain counterpart, the acoustic room impulse response, and the listener's head-related transfer function (HRTF) [1]. Spatial room impulse responses (SRIRs) contain the linear, time-invariant, directional properties of an acoustic environment for a single source-receiver pair and can, together with a set of HRTFs, be used to virtually add a sound object to a real acoustic environment [2, 3]. We thus refer to an SRIR as a multichannel room impulse response that is captured using a somewhat compact microphone array whose inter-element spacing is smaller than a wavelength for a considerable part of the frequency range of interest to pick up directional properties of the room transfer function for a single listener position [4]. SRIRs are often represented in a linear transform domain using spherical harmonics (SH) basis functions which facilitates array-independent analysis and processing [5].

The SRIR of a user's environment is typically unknown and needs to be blindly estimated from signals that naturally occur in that environment. Blind multichannel system identification methods typically exploit cross-relations between the channels to estimate relative transfer functions [6, 7, 8, 9, 10]. While these well-established methods return accurate estimates for short impulse responses, we show in Sec. 3 that they do not converge in reasonable amounts of time in our scenarios of interest, which cover SRIRs of hundreds of milliseconds in length at a sampling rate of $48\,\mathrm{kHz}$. Recently, machine-learning (ML) methods were introduced that either allow for the estimation of room impulse responses or directly match an audio signal with the acoustics of a target environment [11, 12, 13]. However, only [12] achieves this at a sample rate of $48\,\mathrm{kHz}$, supporting the entire audible frequency range. The ML methods so far are designed for single-channel audio signals and do not guarantee the preservation of directional properties of the acoustic environments that are embedded in the inter-channel relations of multichannel SRIRs. As the ML methods need to be trained on large sets of realistic data, training ML models for specific microphone array configurations further requires a large measurement or simulation effort, and methods that do not rely on training data will remain relevant.

In this contribution, we propose a non-data-driven method that uses a pseudo reference signal to transform the blind identification task into a non-blind one and estimate SRIRs using a frequency-domain multichannel Wiener filter. The pseudo reference signal is obtained from the array signals via dereverberation and beamforming. A related method was recently proposed in [14], where relative transfer function estimates in the SH domain were obtained using a beamformer and a frequency-domain recursive-least-squares (RLS) algorithm. The method was evaluated for simulated SRIRs and an ideal SH receiver in terms of a reverberation time error and a directional error. We model a room impulse response as being composed of direct sound, early reflections, and late reverberation, and show that (non-relative) transfer functions can be estimated by a careful design of the beamformer and additional dereverberation. The proposed method directly processes the microphone signals and is not limited to specific microphone arrays but can also be applied to spherical arrays and an SH decomposition of their microphone signals. We systematically investigate the performance of the proposed method in comparison to a generalized version of [14] by analyzing reproduced reverberation times, direct-to-reverberant energy ratios,

[*]We thank Reality Labs Research for funding this research.

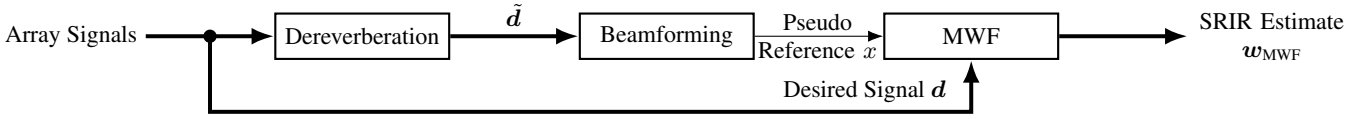

Fig. 1: The SRIR is estimated using a multichannel Wiener filter (MWF) with the multichannel array signals as desired signals and a dereverberated, beamformed signal as pseudo reference signal. Bold lines represent multichannel signals.

and a directional energy metric using 200 simulated and 16 measured SRIRs that include circular microphone arrays, equatorial microphone arrays, and spherical microphone arrays.

2. SPATIAL ROOM IMPULSE RESPONSE ESTIMATION

Let us consider a sound source in a room emitting the signal $s(n)$ which propagates through the room and is being picked up by a microphone array. Under linear, time-invariant conditions, the array signals $\boldsymbol{d}(n)$ are described as a convolution of the signal and the multichannel impulse response $\boldsymbol{h}(n)$, $\boldsymbol{d}(n) = \sum_{l=0}^{L-1} \boldsymbol{h}(l)s(n-l)$. Acoustic room impulse responses are typically described as being composed of three parts, the direct sound $\boldsymbol{h}_{\mathrm{d}}(n)$, the early reflections $\boldsymbol{h}_{\mathrm{e}}(n)$, and the late reverberation $\boldsymbol{h}_{\mathrm{l}}(n)$, so that the convolution can be re-expressed to show the contribution of these individual parts:

$$\begin{aligned}\boldsymbol{d}(n) = &\sum_{l_1=0}^{L_{\mathrm{d}}-1} \boldsymbol{h}_{\mathrm{d}}(l_1)s(n-l_1) + \sum_{l_2=L_{\mathrm{d}}}^{L_{\mathrm{e}}-1} \boldsymbol{h}_{\mathrm{e}}(l_2)s(n-l_2) \\ &+ \sum_{l_3=L_{\mathrm{e}}}^{L-1} \boldsymbol{h}_{\mathrm{l}}(l_3)s(n-l_3)\,. \end{aligned} \tag{1}$$

Blind RIR estimation methods estimate $\boldsymbol{h}(n)$ without access to the source signal $s(n)$. The herein proposed method achieves this via an estimate of $s(n)$, termed pseudo reference signal $x(n)$, that is obtained from the array signals $\boldsymbol{d}(n)$ by dereverberation and beamforming as shown in Fig 1. While the dereverberation aims at canceling the late reverberation in the third term of (1), the beamformer's task is to suppress the second term of (1) containing the influence of the early reflections, while compensating for the direct part of the room impulse response $\boldsymbol{h}_{\mathrm{d}}(n)$ to obtain an undistorted estimate of $s(n)$.

For the dereverberation, we utilize the generalized weighted prediction error (GWPE) method as it offers blind multichannel dereverberation while preserving time differences between channels [15]. In a nutshell, the GWPE method estimates a multichannel linear prediction filter that minimizes temporal signal correlations after a prediction delay. The prediction delay is typically chosen in the range of tens of milliseconds to not cancel the non-zero short-term autocorrelation of dry speech. The GWPE method is usually applied in a subband domain such as the short-term Fourier transform (STFT) domain. We formulate the rest of the processing in the frequency domain and apply the processing to signal blocks obtained via the STFT.

From the dereverberated array signals $\tilde{\boldsymbol{d}}(\omega)$, the pseudo reference signal

$$x(\omega) = \boldsymbol{w}_{\mathrm{BF}}^{\mathrm{H}}(\omega)\tilde{\boldsymbol{d}}(\omega) \tag{2}$$

is obtained via the matched-filter beamformer

$$\boldsymbol{w}_{\mathrm{BF}}(\omega) = \frac{\boldsymbol{a}(\omega)}{\boldsymbol{a}^{\mathrm{H}}(\omega)\boldsymbol{a}(\omega)}\,, \tag{3}$$

where $\boldsymbol{a}(\omega)$ is the array transfer function for a plane wave impinging on the array from the source direction under anechoic conditions, the superscript $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose and ω is the angular frequency. All variables in the remainder of this section are defined in the frequency domain and we omit the frequency dependency for readability. The array transfer function $\boldsymbol{a}$ is assumed to be known for any given source direction and is in practice determined either via an analytic description or measurements. The direction of arrival of the speech signal can, for instance, be estimated via the multiple signal classification (MUSIC) algorithm [16]. If the direction-of-arrival estimate is accurate, the selected anechoic transfer function is equal to the direct part of the RIR, $\boldsymbol{a} = \boldsymbol{h}_{\mathrm{d}}$, and the beamformer recovers the signal s from the direct part of the RIR without distortion. We assume no knowledge about the transfer functions of individual early reflections $\boldsymbol{h}_{\mathrm{e}}$ so that their influence is only suppressed to a certain degree by the directivity of the beamformer. The matched-filter beamformer can be interpreted as a minimum-variance distortionless response (MVDR) beamformer with the identity matrix as the noise power spectral density (PSD) matrix, which for instance is the case for spherical arrays with regularly distributed microphones in an isotropic white-noise field.

A frequency-domain multichannel Wiener filter (MWF) [17, Ch. 6.6] is used to estimate the SRIR. It is designed to minimize the mean square error (MSE) between the filtered pseudo reference signal $\boldsymbol{y} = \boldsymbol{w}x$ and the array signal $\boldsymbol{d}$,

$$\begin{aligned}\boldsymbol{J}_{\mathrm{MSE}}(\boldsymbol{w}) &= \mathrm{E}\{\|\boldsymbol{y} - \boldsymbol{d}\|_2^2\}\,, \\ &= \mathrm{E}\{x^*\boldsymbol{w}^{\mathrm{H}}\boldsymbol{w}x\} - 2\mathrm{E}\{x^*\boldsymbol{w}^{\mathrm{H}}\boldsymbol{d}\} + \mathrm{E}\{\boldsymbol{d}^{\mathrm{H}}\boldsymbol{d}\}\,. \end{aligned} \tag{4}$$

The superscript $(\cdot)^*$ denotes complex conjugation and $\mathrm{E}\{\cdot\}$ is the expectation. The optimal filter $\boldsymbol{w}_{\mathrm{MWF}}$ is obtained after setting the derivative $\partial\boldsymbol{J}_{\mathrm{MSE}}(\boldsymbol{w})/\partial\boldsymbol{w}$ to zero,

$$\boldsymbol{w}_{\mathrm{MWF}} = \frac{1}{\Phi_{xx}}\boldsymbol{\Phi}_{xd}\,, \tag{5}$$

where $\Phi_{xx} = \mathrm{E}\{x^*x\}$ is the PSD of the pseudo reference signal and $\boldsymbol{\Phi}_{xd} = \mathrm{E}\{x^*\boldsymbol{d}\}$ is the cross spectral density (CSD) vector of the pseudo reference and the desired signal. In this publication, we apply the processing directly to the array signals but as there are no specific assumptions on the microphone array, the whole processing can be performed in any linear transform domain such as the spherical harmonics domain [5]. We consider batch processing of a few seconds of captured array signals but adaptive dereverberation algorithms based on multichannel linear prediction are available [18, 19], and the MWF can be replaced by a recursive-least-squares algorithm [14].

3. CONVERGENCE BEHAVIOR

Acoustic environments typically have reverberation times of several hundreds of milliseconds to multiple seconds. At sampling rates that support the full audible frequency range up to 20 kHz, this results

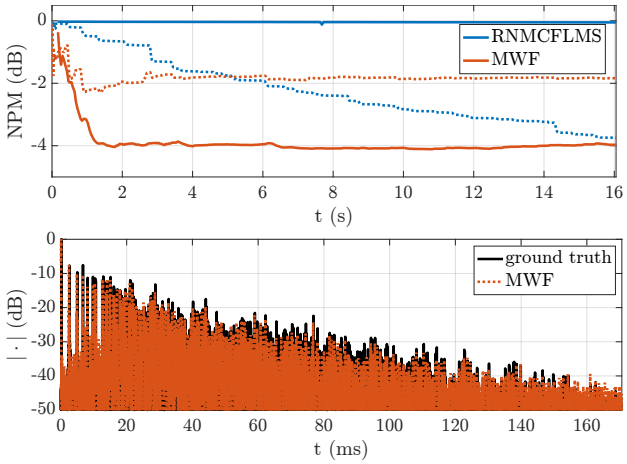

Fig. 2: Top: NPM for RNMCFLMS and MWF (proposed), for SRIR lengths of 512 samples (dotted lines) and 8192 samples (solid lines). Bottom: Magnitude of one channel of the 8192-sample ground truth SRIR and the estimate from the MWF.

in lengths of the corresponding room impulse responses of tens of thousands to hundreds of thousands of samples. To successfully identify such long responses from a few seconds of speech, fast convergence is required. We analyze the convergence behavior of the proposed method by calculating the normalized projection misalignment (NPM) [20] of SRIR estimates from the proposed method and comparing the convergence behavior to the cross-relation-based Robust Normalized Multichannel Frequency-Domain Least Mean Square (RNMCFLMS) algorithm [9]. The NPM is a scale-independent measure of the norm of the error between an estimated impulse response and the ground truth impulse response. Fig. 2 (top) shows the NPM over time that is achieved by the proposed method (MWF) and the RNMCFLMS for an SRIR that was simulated for a spherical microphone array with 6 microphones using the image source method and a sampling rate of $48\,\mathrm{kHz}$. While the RNMCFLMS algorithm achieves a lower NPM than the proposed solution for the shortened SRIR of 512 samples (dotted lines), the RNMCFLMS method does not show a meaningful reduction of the NPM when estimating an 8192-sample-long SRIR (solid lines) which corresponds to a length of $170\,\mathrm{ms}$ and thus is still rather short for a realistic acoustic impulse response. The proposed method in both cases converges within less than 2 seconds.

The lower part of Fig. 2 shows one channel of the 8192-sample ground truth SRIR and the corresponding estimate of the proposed method. As suggested by the corresponding NPM of about $-4\,\mathrm{dB}$, the proposed method is not able to accurately reproduce all of the individual reflection peaks but, as illustrated by the shown magnitude, captures overall characteristics like the decay rate and the reflection density well. We thus evaluate the performance of the proposed method using three different room acoustical metrics in the following. Due to the poor convergence behavior of the RNMCFLMS even with SRIRs of moderate length, we exclude it from the evaluation.

4. EVALUATION

4.1. Evaluation Procedure

We evaluate the proposed method using three metrics and compare it to a generalized version of [14] that does not require an SH decomposition and is obtained by removing the dereverberation from the proposed processing shown in Fig. 1. For comparability, we do not use the adaptive solution from [14] but also utilize the MWF in this dereverberation-free case which is equivalent to running the method from [14] with a forgetting factor of one. The evaluation scenarios are static so that an adaptive solution is not required.

The chosen evaluation metrics comprise the reverberation time T_{20}, the direct-to-reverberant energy ratio (DRR), and the weighted angular error [14]. While the first two metrics are individually obtained for each microphone channel, the latter characterizes the directional energy distribution that is captured by the inter-channel relations of the SRIRs. The T_{20} error ϵ_{T20} is calculated as the absolute difference of the T_{20} reverberation time in octave bands of the estimated SRIR and the ground truth, and is given in percent relative to the ground truth. It is calculated using the toolbox from [21]. The DRR error ϵ_{DRR} is obtained as the energy ratio of the direct sound (in a window starting $1\,\mathrm{ms}$ before and ending $2\,\mathrm{ms}$ after the direct sound peak) and the rest of the SRIR until the beginning of the noise floor which is determined by the T_{20} estimator. The weighted angular error ϵ_{ang} is calculated by representing the array signals via spherical harmonic (SH) or circular harmonic (CH) coefficients using the least-squares approach from [22] and calculating the angular mismatch of the pseudo intensity vector (PIV) between the estimated SRIR and the ground truth as done in [14]. We use a decomposition into SHs for arrays with microphones distributed on a spherical surface and CHs for arrays with microphones on a circle. The error is calculated for the first $50\,\mathrm{ms}$ of the SRIRs and the final results are given in degrees. All metrics are obtained for a frequency range between $200\,\mathrm{Hz}$ and $8\,\mathrm{kHz}$ where the employed speech signals have significant energy. Although accurate results for the given metrics are only expected in this frequency range, the estimates are generated at a sampling rate of $48\,\mathrm{kHz}$ to support a potential auralization.

The error metrics were evaluated for a total of 200 simulated and 16 measured SRIRs. SRIRs of $200\,\mathrm{ms}$ length were simulated for 20 arrays and the same 10 rooms per array using the image source method with the tool from [23]. The simulated shoebox-shaped rooms had random dimensions between $4 \times 4 \times 2$ m and $10 \times 8 \times 5$ m generated from an equal distribution. The source and the microphone array were placed at random positions in the rooms while ensuring a minimum distance of $1\,\mathrm{m}$ to the walls and $2\,\mathrm{m}$ to each other. The same random absorption coefficients with equally distributed values between 0.2 and 0.7 were applied to all boundaries, resulting in reverberation times between $240\,\mathrm{ms}$ and $480\,\mathrm{ms}$. The simulated arrays comprise open and rigid spherical microphone arrays (SMAs), i.e., microphone arrays with microphones that are distributed on the surface of a sphere according to t-designs [24] with and without a spherical scattering body, circular microphone arrays (CMAs), and equatorial microphone arrays (EMAs) [25]. For CMAs and EMAs, the microphones are equally distributed on a circle, and for EMAs, this circle is located on the equator of a rigid sphere. The arrays were simulated with radii of $4\,\mathrm{cm}$ and $8\,\mathrm{cm}$, and with 6, 12, and, in the case of the SMAs, 32 microphones.

The measurement-based evaluation was performed with openly accessible SRIRs from a variable acoustics room [26]. In particular, we used 4 measurement positions of the Eigenmike EM32 32-channel, rigid-sphere microphone array with a radius of $4.2\,\mathrm{cm}$ and 4 measurement positions of the Zylia ZM-1 19-channel, rigid-sphere array with a radius of $4.9\,\mathrm{cm}$. For both arrays and all measurement positions, we compared measurements with $25\,\%$ and $50\,\%$ active absorption in the room (leading to reverberation times of $760\,\mathrm{ms}$ and $540\,\mathrm{ms}$ at $1\,\mathrm{kHz}$), resulting in a total of 16 SRIRs.

The SRIRs were estimated using two different speech signals

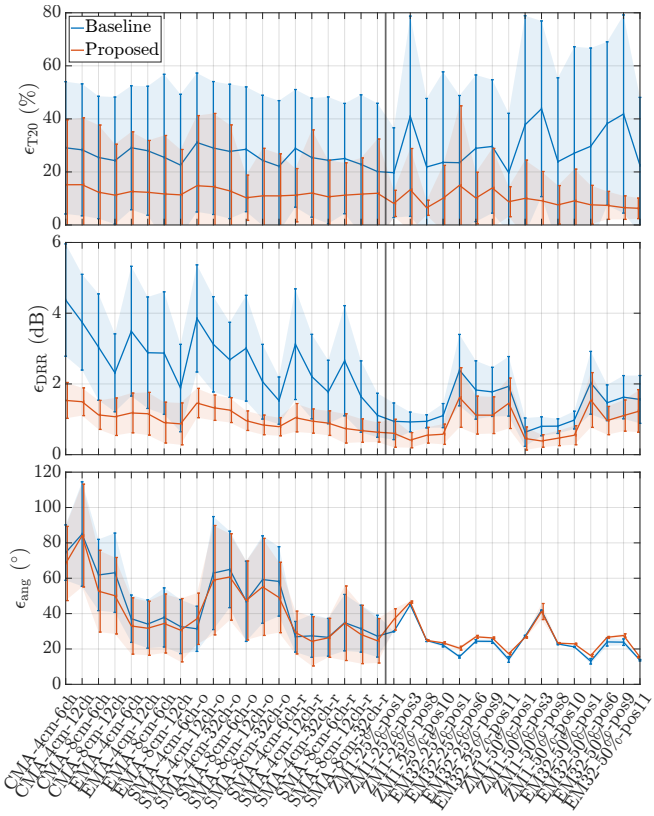

Fig. 3: Means and standard deviations of the reverberation time error ϵ_{T20}, the DRR error ϵ_{DRR}, and the weighted angular error ϵ_{ang}. Results from the simulated and the measured SRIRs are separated by a vertical black line.

from the EBU SQAM[1], one containing male and one female speech, which were convolved with the ground truth SRIR to obtain the array signals that were used for the SRIR estimation. The array signals were created at a sampling rate of 48 kHz and had a length of 2.5 seconds (male speech) and 4 seconds (female speech). The GWPE dereverberation was applied in blocks of 2048 samples, with a hop size of 128 samples, a prediction order of 12, and a prediction delay of 20 ms. In the case of the simulated SRIRs, the beamformer was informed of the true direction of arrival (DOA) of the speech, and in the case of the measured SRIRs, a broadband DOA was estimated using the MUSIC algorithm [16]. The PSDs and CSDs of the MWF were estimated in signal blocks of 400 ms length for the simulated SRIRs and 1 s for the measured SRIRs, which is significantly longer than the true responses to facilitate the approximation of the convolutive transfer functions as multiplicative transfer functions via the STFT, and with a hop size of 43 ms.

4.2. Results

Fig. 3 shows the means and standard deviations of the three error metrics which were calculated from the individual errors for each of the two speech signals, each microphone channel (only for ϵ_{T20} and ϵ_{DRR}), and each octave band (only ϵ_{T20}). In the case of the simulated SRIRs (shown left of the vertical black line in the figure), results were further averaged over 10 different rooms for each array which explains the higher standard deviations when compared to the results from the measured SRIRs. The axis labels for the simulated SRIRs denote the type of microphone array, the array radius, and the number of microphone channels, while the labels for the measured SRIRs denote the array type, the percentage of active absorption in the room, and the measurement position.

The proposed method achieves mean reverberation time errors ϵ_{T20} between 6.3 % and 15.2 % which are in all cases lower than the mean errors of the baseline method without the dereverberation ranging between 19.7 % and 43.9 %. For the measured SRIRs, the proposed method further shows significantly lower standard deviations than the baseline method. The type of microphone array influences the ϵ_{T20} performance only slightly but in many cases, arrays with a larger radius show slightly lower mean ϵ_{T20} errors than their counterpart with a smaller radius.

Similar observations hold for the DRR error ϵ_{DRR}. Again, the proposed method in all cases generates lower mean errors than the baseline and also smaller standard deviations, especially for the simulated SRIRs. The mean ϵ_{DRR} errors of the proposed method lie in a range between 0.4 dB and 1.6 dB while the method without the dereverberation produces mean errors between 0.6 dB and 4.4 dB. Again, arrays with a larger radius show better performance.

In the case of the weighted angular error ϵ_{ang}, both methods perform similarly: the proposed method achieves angular errors between 15° and 84°, and the baseline method between 13° and 85°. As seen from the results for the simulated arrays, arrays with a rigid scattering body achieve considerably lower angular errors than arrays without a scattering body. This is not surprising as the rigid spherical scattering body is known to facilitate a better conditioned SH representation [5, Ch. 4.6].

For the target application of auditory augmented reality, the goal is to create perceptually convincing virtual sound sources from the SRIR estimates. Just noticeable differences (JNDs) for the DRR are known to be between 2.4 dB and 7.3 dB depending on the DRR [27]. By interpolating the given JNDs from [27] that were determined for DRRs of -10, 0, 10, and 20 dB, we found that 98.9 % of the estimated DRRs of the proposed method deviate from the ground truth by less than the JND. JNDs for the reverberation time lie between 4 % and 7 % [28, 29] and thus the estimates of the proposed method in most cases must be assumed to be perceptually distinguishable from the ground truth. However, perceivable differences do not necessarily impair the plausibility of renderings, and the results may thus still be valuable for the target application of auditory augmented reality applications where virtual sound sources may differ from real sources in signal and position and are only indirectly comparable.

5. CONCLUSION

We presented a method for the blind estimation of spatial room impulse responses (SRIRs) from speech signals using dereverberation and a beamformer. The method is able to estimate SRIRs of realistic lengths using just a few seconds of speech signals. The SRIR estimates reproduce the reverberation time with an average error below 16 % for all arrays and 98.9 % of the estimates reproduce the direct-to-reverberant energy ratio with errors below the just noticeable difference. A perceptual evaluation of the estimated SRIRs is planned as future work.

[1] Sound Quality Assessment Material recordings for subjective tests, available at `https://tech.ebu.ch/publications/sqamcd`.

6. REFERENCES

[1] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Riitta Väänänen, "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, 1999.

[2] Leo McCormack, Nils Meyer-Kahlen, and Archontis Politis, "Multi-directional parameterisation and rendering of spatial room impulse responses," in *Proc. of the 24th International Congress on Acoustics*, 2022.

[3] Thomas Deppisch, Sebastià V. Amengual Garí, Paul Calamia, and Jens Ahrens, "Perceptual Evaluation of Spatial Room Impulse Response Extrapolation by Direct and Residual Subspace Decomposition," in *AES International Conference on Audio for Virtual and Augmented Reality*, 2022, p. 1–10.

[4] Thomas Deppisch, Sebastià V. Amengual Garí, Paul Calamia, and Jens Ahrens, "Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 927–942, 2023.

[5] Boaz Rafaely, *Fundamentals of Spherical Array Processing*, Springer, 2nd edition, 2019.

[6] Guanghan Xu, Hui Liu, Lang Tong, and Thomas Kailath, "A Least-Squares Approach to Blind Channel Identification," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[7] Yiteng Huang and Jacob Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, 2003.

[8] Sharon Gannot and Marc Moonen, "Subspace Methods for Multimicrophone Speech Dereverberation," *EURASIP J. on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[9] Mohammad Ariful Haque and Md Kamrul Hasan, "Noise Robust Multichannel Frequency-Domain LMS Algorithms for Blind Channel Identification," *IEEE Signal Processing Letters*, vol. 15, pp. 305–308, 2008.

[10] Byeongho Jo and Paul Calamia, "Robust blind multichannel identification based on a phase constraint and different lp-norm constraints," in *28th European Signal Processing Conference*, 2021, pp. 1966–1970.

[11] Jiaqi Su, Zeyu Jin, and Adam Finkelstein, "Acoustic Matching By Embedding Impulse Responses," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 426–430.

[12] Christian J. Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia, "Filtered Noise Shaping for Time Domain Room Impulse Response Estimation from Reverberant Speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2021, pp. 221–225.

[13] Anton Ratnarajah, Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, Pablo Hoffmann, Dinesh Manocha, and Paul Calamia, "Towards Improved Room Impulse Response Estimation for Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[14] Nils Meyer-Kahlen and Sebastian J Schlecht, "Blind Directional Room Impulse Response Parameterization from Relative Transfer Functions," in *Proc. Int. Workshop on Acoustic Signal Enhancement*, 2022, p. 1–5.

[15] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

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

[17] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing*, Springer Berlin, Heidelberg, 2008.

[18] Takuya Yoshioka and Tomohiro Nakatani, "Dereverberation for reverberation-robust microphone arrays," in *21st European Signal Processing Conference*, 2013, pp. 1–5.

[19] Sebastian Braun and Emanuel A.P. Habets, "Online Dereverberation for Dynamic Scenarios Using a Kalman Filter with an Autoregressive Model," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1741–1745, 2016.

[20] Dennis R. Morgan, Jacob Benesty, and M. Mohan Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Processing Letters*, vol. 5, no. 7, pp. 174–176, 1998.

[21] Marco Berzborn, Ramona Bomhardt, Johannes Klein, Jan-Gerrit Richter, and Michael Vorländer, "The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing," *Proc. of the German Annual Conference on Acoustics (DAGA)*, pp. 222–225, 2017.

[22] Sébastien Moreau, Jérôme Daniel, and Stéphanie Bertet, "3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of a 4th Order Spherical Microphone," in *120th Conv. Audio Eng. Soc.*, 2006.

[23] Daniel P. Jarrett, Emanuël A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, 2012.

[24] R. H. Hardin and N. J.A. Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions," *Discrete and Computational Geometry*, vol. 15, no. 4, pp. 429–441, 1996.

[25] Jens Ahrens, Hannes Helmholz, David Lou Alon, and Sebastià V. Amengual Garí, "Spherical harmonic decomposition of a sound field based on observations along the equator of a rigid spherical scatterer," *The Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 805–815, 2021.

[26] Thomas McKenzie, Leo McCormack, and Christoph Hold, "Dataset of Spatial Room Impulse Responses in a Variable Acoustics Room for Six Degrees-of-Freedom Rendering and Analysis," *arXiv:2111.11882*, pp. 1–3, 2021.

[27] Erik Larsen, Nandini Iyer, Charissa R. Lansing, and Albert S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 7 2008.

[28] H.P. Seraphim, "Untersuchungen über die Unterschiedsschwelle exponentiellen Abklingens von Rauschbandimpulsen," *Acta Acustica united with Acustica*, vol. 8, no. 4, pp. 280–284, 1958.

[29] Matti Karjalainen and Hanna Jarvelainen, "More about this reverberation science: Perceptually good late reverberation," *Proc. 111th Conv. Audio Eng. Soc.*, pp. 1–8, 2001.