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

On the detection quality of early room reflection directions using compressive sensing on rigid spherical microphone array data

Frank SCHULTZ, Sascha SPORS

Institute of Communications Engineering, University of Rostock, Germany, {frank.schultz, sascha.spors}@uni-rostock.de

Abstract

The estimation of acoustic reflection coefficients from in-room measurements using one or more sources and microphone arrays has been addressed in various ways. One particularly illustrative method is the plane wave decomposition. Peaks in this representation can be assigned to the location and strength of mirror image sources. Traditional approaches employ rigid spherical microphone arrays together with modal beamforming. Compressive sensing (CS) techniques aim at spatial undersampling by assuming sparsity of the sound field in some given representation. By careful design of the sensing matrix, the number of required measurements can be considerably reduced compared to Nyquist sampling and reconstruction.

However, often sparsity of the problem at hand is not given and a sufficient low mutual coherence of the sensing matrix is violated, yielding CS reconstruction with low robustness. We aim at robust detection of early, discrete reflections by means of CS using rigid spherical microphone array data. We discuss the interaction of room characteristics, sensing matrix and technical measures to qualify the detection performance. The influence of practical limitations, like the number of microphones and their self-noise is investigated, as well as the overall gain of applying CS to the problem compared to traditional modal beamforming techniques.

Keywords: early room reflections, microphone array, sound field analysis, modal beamforming, compressive sensing

1 INTRODUCTION

Direction of arrival (DOA) estimation using sensor arrays is a well addressed problem in array processing, for which different approaches with advantages and disadvantages exist [1]. Spherical microphone arrays allow for DOA estimation by beamforming in the modal domain using spherical harmonics (SHs). With the advent of high computational power, the modal plane wave decomposition (PWD) with specific weights (such as matched filter characteristics, delay-and-sum, minimum variance distortionless response) became a useful tool for rigid and open sphere microphone array measurements featuring full audio bandwidth [2]. Compressive sensing (CS) techniques [3] have been increasingly utilized for acoustic problems in the last decade [4]. CS allows for highly undersampling the signals with – in the ideal case – perfect reconstruction, but only if sparsity in a certain (unknown and/or unaccessible) signal domain holds. Naturally, CS was soon applied to acoustic beamforming and DOA estimation, e.g. [5, 6], in the presented scenarios outperforming the aforementioned approaches. Reconstruction of reverberant sound fields by CS was pursued e.g. in [7, 8], indicating very good reconstruction for the very early part of a sound field and for very low audio frequencies.

In this contribution, we aim at reliable DOA estimation of early room reflections [9], i.e. peaks of the room impulse response arriving considerably earlier than the so called mixing time [10] of the room. We compare a traditional PWD approach against one based on CS. Since rigid spherical microphone arrays in range of 4-15 cm radius are nowadays available for practical in-situ room measurements, we concentrate on these designs, rather than randomly sampled or larger aperture microphone arrays. The question arises if CS can be profitably deployed for DOA estimation under these circumstances rather than the PWD approach.

2 METHOD AND SETUP

We approach the evaluation by a specific example, mainly following the study [9], where the DOAs of room reflections and their respective wall absorption coefficients were estimated by a PWD. The room under discussion has a size of (width$_\mathrm{x}$, length$_\mathrm{y}$, height$_\mathrm{z}$) = (10, 11, 3) m. The source is located at $\mathbf{x}_\mathrm{S} = (7, 8, 1.8)^\mathrm{T}$m, where $\mathbf{x} = (x, y, z)^\mathrm{T}$ denotes a vector in Cartesian coordinates. The receiver position is $\mathbf{x}_\mathrm{R} = (2.1, 4.5, 1.8)^\mathrm{T}$m. The room walls shall be built from brick, i.e. a frequency independent absorption coefficient of 0.15 is assumed. The simulations use a sampling frequency of $f_\mathrm{s} = 16\,\mathrm{kHz}$ and a speed of sound of $\mathrm{c} = 343\,\mathrm{m/s}$. The perceptual mixing time is predicted [10] to $t_{mp,50\%} = 31.2\mathrm{ms}$ and $t_{mp,95\%} = 54\mathrm{ms}$. Thus, we restrict our discussion to the time window $0 \leq t \leq 40\,\mathrm{ms}$, where the assumption of 'incidence angle over time'-sparsity for the reflections should be valid. For this time window the direct sound and 31 reflections from up to 4th order mirror image sources arrive at the microphone array, cf. the detailed list in Figure 5. Only for the simplified and introductionary scenario discussed in Figure 1, a 1st (instead of 4th) order image source model with fully absorbing floor and ceiling (instead of brick) is used, while maintaining all other parameters described above. In the acoustic impulse response this results in 1st order reflections in the horizontal plane only.

The temporal domain acoustic transfer matrix $\mathbf{A}_{N\times t\times M}$ (in CS usually termed sensing matrix) between plane wave incidence and microphone array receivers is built from fixed spherical Lebedev grid's $M = 1202$ angle pairs (azimuth ϕ_m, colatitude θ_m) of plane wave incidence to $N = 14, 26, 38, \mathrm{or}\, 170$ microphones arranged on a spherical Lebedev grid $(\phi_\mathrm{n}, \theta_\mathrm{n})$ with radius $R_\mathrm{N} = 0.1\,\mathrm{m}$ centered at position $\mathbf{x}_\mathrm{R}$. As stated above, we do not deploy random sampling for the microphone array positions, since we are rather interested in using already built rigid sphere microphone arrays. In consequence, using the Lebedev grid both for receiver as well as for the DOA dictionary, leads to a rather high mutual coherence. Note that monte carlo simulations for plane wave incidence on rigid spherical microphone arrays (not discussed here) indicate that random sampling does not necessarily decrease mutual coherence compared to regular sampling schemes.

The measurement matrix $\mathbf{y}_{N\times t}$ of room reflections from $(\phi_\mathrm{is}, \theta_\mathrm{is})$ impinging onto $(\phi_\mathrm{n}, \theta_\mathrm{n})$ of the rigid microphone array is obtained by the image source model, being implemented in the SH and temporal frequency domain using sufficient temporal resolution. Inverse spatial and temporal Fourier transforms yield the acoustic impulse responses $\mathbf{y}_{N\times t}$ for the N microphones. Uncorrelated white Gaussian noise with dedicated signal peak to noise variance ratio is (if any) then added to simulate sensor noise. Note that this simple additive noise model holds only for impulse response measurements by impulsive excitation, cf. [11].

DOA estimation via the plane wave decomposition includes matched filtering in the SH domain, yielding a high frequency independent white noise gain, cf. [9] for details. The PWD matrices $\mathbf{p}[N]_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}$ for the $N = 14, 26, 38, 170$ microphone grids and their conservatively chosen corresponding SH orders $2, 3, 4, 10$ are obtained by inverse SH transform for all plane wave dictionary entries $(\phi_\mathrm{m}, \theta_\mathrm{m})$. Typically, either one angular dependency over time is conveniently visualized in a surface plot, cf. [9, Fig. 6], or θ over ϕ is plotted for a fixed time instance, cf. [2, Fig. 5.6].

DOA estimation by means of compressive sensing uses orthogonal matching pursuit (OMP) [3, Ch. 4] in its implementation in Python's scikit-learn package. A sliding window with length $t_w = 0.75\,\mathrm{ms}$ (according to a chosen sound path of $2.5725 R_\mathrm{N}$, i.e. 12 samples at f_s) for all considered t is utilized to solve the system of equations $\mathbf{y}_{N\cdot t_w\times 1} = \mathbf{A}_{N\cdot t_w\times M} \cdot \mathbf{x}_{M\times 1}$ (matrix/vector stacking) for $\mathbf{x}_{M\times 1}$ with OMP. We search for 3 non-zero coefficients in $\mathbf{x}_{M\times 1}$ per time frame. This seems to be arbitrary, but is linked to the temporal coincidence of certain reflections. Thus, this choice is highly dependent on the room characteristics. Then, finding a specific OMP solution for a specific t, the sparse matrix $\mathbf{x}_{M\times t}$ with $3M$ non-zero coefficients is obtained. Explicitly rearranging for plane wave incidence angles and considering the different microphone array grids N gives the DOA matrices $\mathbf{x}[N]_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}$ using CS-OMP beamforming.

Feel free to use the paper's dedicated Jupyter notebook[1] to conduct own simulations.

[1] `https://github.com/spatialaudio/doa-early-room-reflections-pwd-vs-omp`

3 RESULTS AND DISCUSSION

In the following we aim at direct comparison of both, PWD and CS-based DOA matrices $\mathbf{p}[N]_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}$ and $\mathbf{x}[N]_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}$ using the same setups. These raw data could be used to compute DOA estimates by e.g. a blob or peak detection.

For the three-dimensional (3D) DOA results over time, the four dimensions ϕ, θ, t and level of the early reflections are to be visualized for presentation. The following scatter plots seem to be a reasonable way to do so. Note, that normal surface plots can easily provide a misleading picture of 4D data, especially when being sparse. The left scatter plots in the figures show DOA as level over azimuth over time. The right scatter plots show DOA as level over colatitude over time. The top scatter plots show PWD based DOA matrix $|\mathbf{p}_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}|$ in dB. The bottom scatter plots show CS-OMP based DOA matrix $|\mathbf{x}_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}|$ in dB. The level is normalized to 0 dB for the direct sound and exhibits a visible range from -10 dB to 1 dB. Under the assumption that mirror image sources / reflections behave like point sources, their amplitude decay $\propto 1/r$ over distance r can be compensated for. This is simply achieved by estimating $\frac{1}{r} = \frac{1}{ct}$ from the traveling time t starting from the direct sound arrival. The method is actually exact for images sources of 1st order, only. Otherwise, for higher order image sources (multiple wall reflections) the absorption coefficients must be rather low, in order that this compensation is meaningful.

Color and size of the scatter circles indicate level. In the top left corner of each subplot two reference circles are included, not to be mistaken with estimated reflections: The large yellow circle indicates the size for 0 dB

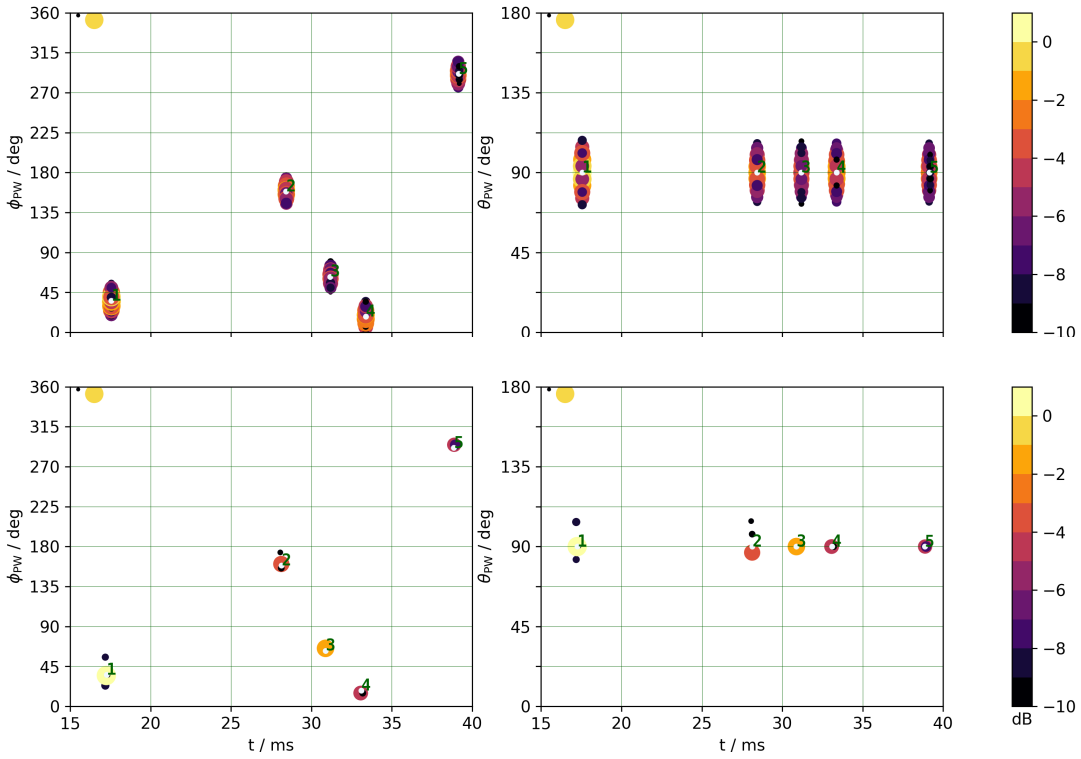

Figure 1. Early room reflections' DOA over time. Left: azimuth, right: colatitude, top: PWD $|\mathbf{p}_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}|$ in dB, bottom: CS-OMP $|\mathbf{x}_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}|$ in dB. Rigid microphone array, spherical Lebedev grid with 170 receivers, 1st order image source model, fully absorbing ceiling and floor, no additive noise.

level, the small black circle -10 dB. Circle's radius in between is varied linearly in terms of dB. A small white dot and a green number right above it indicate the reference position of an image source. Numbering is sorted according to traveling time. Scatter plots could be improved in future by overlaying low to high levels.

Since the degree of freedom for variation of parameters is huge, we restrict the discussion to four prototypical scenarios with distinct features. The results for the first scenario are depicted in Figure 1 for $N = 170$ and 1st order image sources in a room with fully absorbing ceiling and floor (cf. courtyard with snow-covered meadow). This rather artificial scenario links to the setup used in [9] and aims at introducing the visualization strategy. It can be seen, that both, the PWD and the CS-OMP, could reliably detect all five DOAs (direct sound at 17.556 ms due to $|\mathbf{x}_\mathrm{S} - \mathbf{x}_\mathrm{R}| = 6.02\,\mathrm{m}$ as well as the four reflections from the four walls) in the horizontal plane. The PWD exhibits the typical elliptical shapes around level maxima (here i.e. the main lobe) in this visualization, which correspond to circular beam patterns mapped on a sphere. A robust blob detection in searching for level maximum matching the focal point of the ellipses / circles would rely on a clear separation of these shapes (no smearing of grating and side lobes). OMP by design exhibits sparse DOA events. Even in this simple scenario multiple detection of certain DOAs occur, cf. prominent slightly misaligned around events 1 and 2, which however would not provide misleading results as they can clearly distinguished from the maximum level entry at about the same time instance.

The second scenario for Figure 2 uses $N = 170$ receivers, all walls built from brick and image sources up to 4th order. Thus, the DOAs of the direct sound and 31 reflections are to be estimated. Ellipses within PWD can be distinguished clearly by visual inspection up to the 11th mirror image source. The clusters around 33 ms

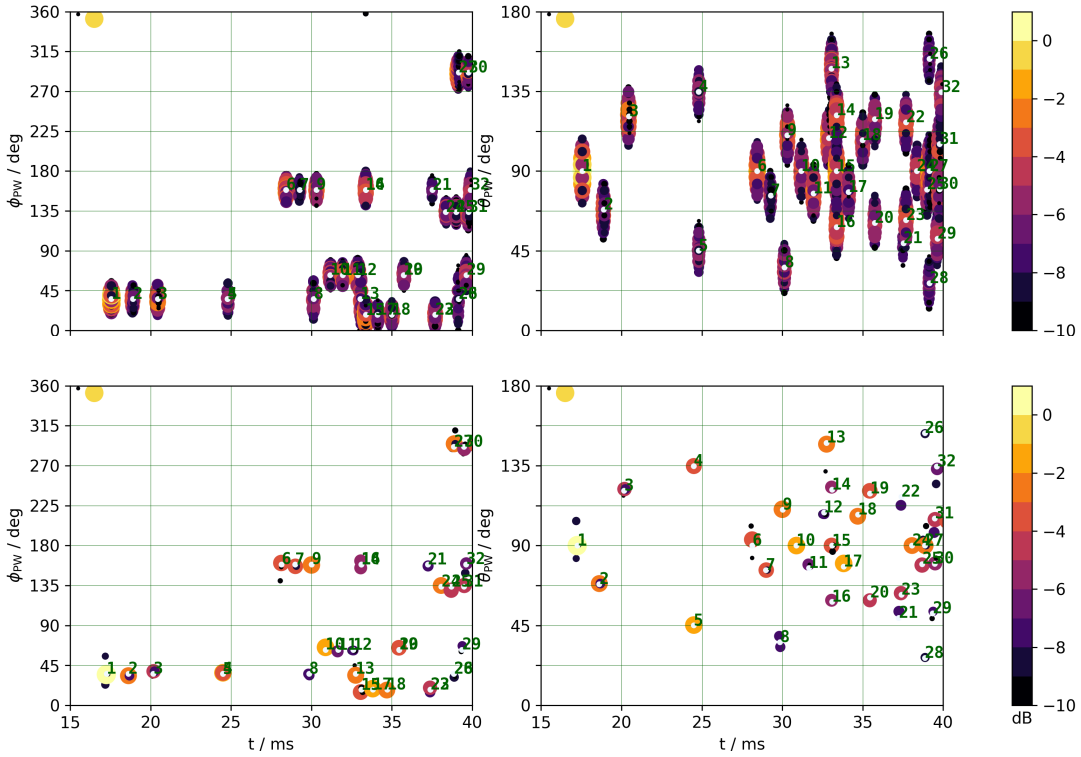

Figure 2. Early room reflections' DOA over time. Left: azimuth, right: colatitude, top: PWD $|\mathbf{p}_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}|$ in dB, bottom: CS-OMP $|\mathbf{x}_{\phi_\mathrm{m}\times\theta_\mathrm{m}\times t}|$ in dB. Rigid microphone array, spherical Lebedev grid with 170 receivers, 4th order image source model, all walls brick (absorption coefficient 0.15), no additive noise.

and around 39 ms are likely not detectable, whereas in between certain reflections are prominently isolated, cf. mirror image source 20. This is again highly dependent on the room characteristics. The CS-OMP based DOA is highly precise *if* the DOA was correctly found, which is here always the case.

The third scenario for Figure 3 uses $N = 170$ receivers, all walls built from brick, image sources up to 4th order and uncorrelated white noise added to the receivers with rather high peak to noise ratio of 10 dB. This is chosen to conveniently discuss the general effect of noisy measurements when comparing our PWD and CS-OMP DOA methods. Recall that we compensate for the $1/r$ amplitude decay, which raises noise with increasing time. This degrades the DOA results in the PWD as well as in the CS-OMP. The characteristics are however different. Visual inspection allows for robust DOA estimation up to the 11th event in the PWD, the same as for the example without noise. These ellipses are not degraded too much by the noise. CS-OMP based DOA becomes blurry for $t > 27.5\,\mathrm{ms}$. Also the first 5 mirror image sources are not indisputable, but uncritical for robust DOA estimation. However, this blur leads to an increasing amount of DOA clustering for increasing time. While PWD estimation would ensure robust estimation of the first 11 mirror image sources, this is not save for CS-OMP in this example, cf. false positives with high levels around events 11, 17 and 18. Conversely, with CS-OMP a robust-like detection of e.g. event 14 in very noisy data would be possible, where PWD data would fail.

The fourth scenario for Figure 4 uses $N = 14$ receivers, all walls built from brick and image sources up to 4th order, no noise. Thus, one prerequisite for CS, $N \ll M$, is best met. Moreover, N is about half as the

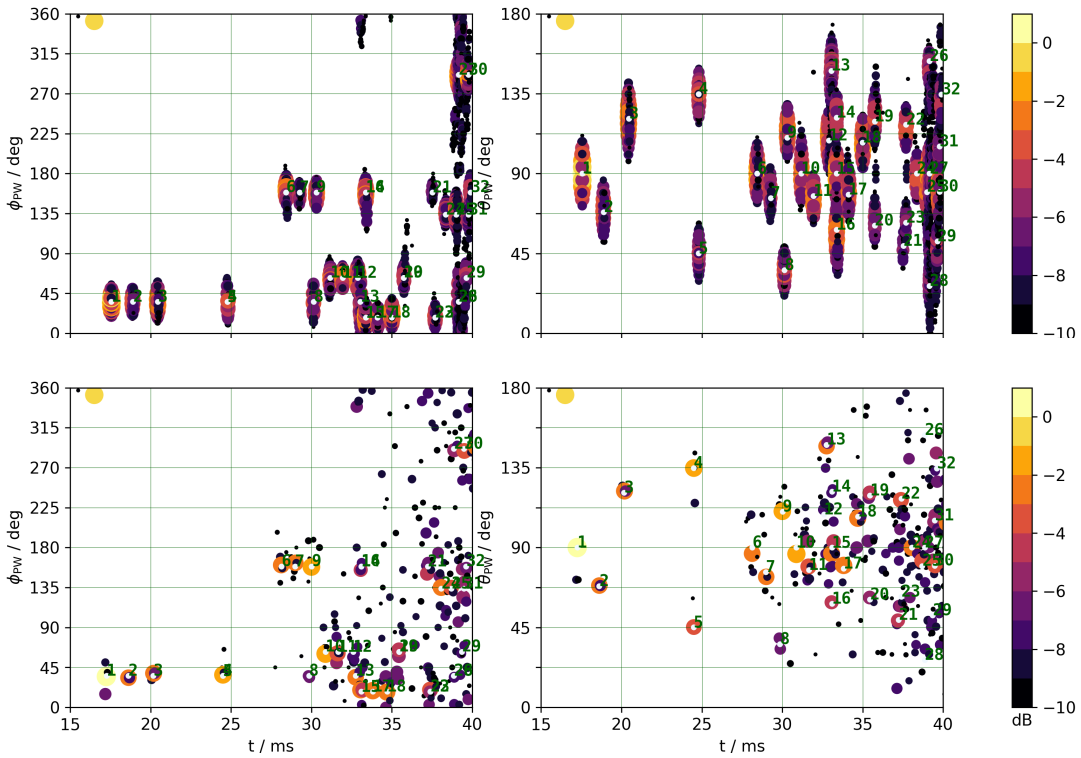

Figure 3. Early room reflections' DOA over time. Left: azimuth, right: colatitude, top: PWD $|\mathbf{p}_{\phi_m \times \theta_m \times t}|$ in dB, bottom: CS-OMP $|\mathbf{x}_{\phi_m \times \theta_m \times t}|$ in dB. Rigid microphone array, spherical Lebedev grid with 170 receivers, 4th order image source model, all walls brick (absorption coefficient 0.15), 10 dB signal peak to noise variance ratio.

number of early reflections to be estimated by DOA, which should be considered as a challenging scenario even without noise. Robust DOA estimation based on the PWD data by visual inspection is meaningful for direct sound and first two reflections. Fourth and fifth events can not be reliably separated. The CS-OMP based DOA is blurry. The blur characteristics are different to the previous example. For $N = 14$, the blur is spread over whole time range due to sparse spatial sampling, whereas for $N = 170$ the blur becomes more prominent for increasing time due to the 1/r compensation and thus increased noise level. Robust detection with CS-OMP would be possible for the events 1 to 9. For all other events, false positives exhibit levels that could not be ignored when assigning early reflections on a blind basis. For example note that, the false positives at 24 ms with considerable level suggest a further reflection besides 4th and 5th mirror image source.

4 CONCLUSION

We presented a study comparing DOA estimation for early room reflections. Since for these, sparsity in the acoustic impulse response can be assumed, it appeared desirable to apply compressive sensing based beamforming. In particular we utilized rigid microphone arrays with spherical Lebedev grids, as these are most practical for fast in-situ measurements. We compared DOA by matched filter plane wave decomposition and by compressive sensing using orthogonal matching pursuit with a sliding window and up to 3 non-zero coefficients per window. A consistent comparison for 3D data over time was missing so far to the authors knowledge. Our simulations and the brief presentation of four special cases indicate that both approaches generally are suitable

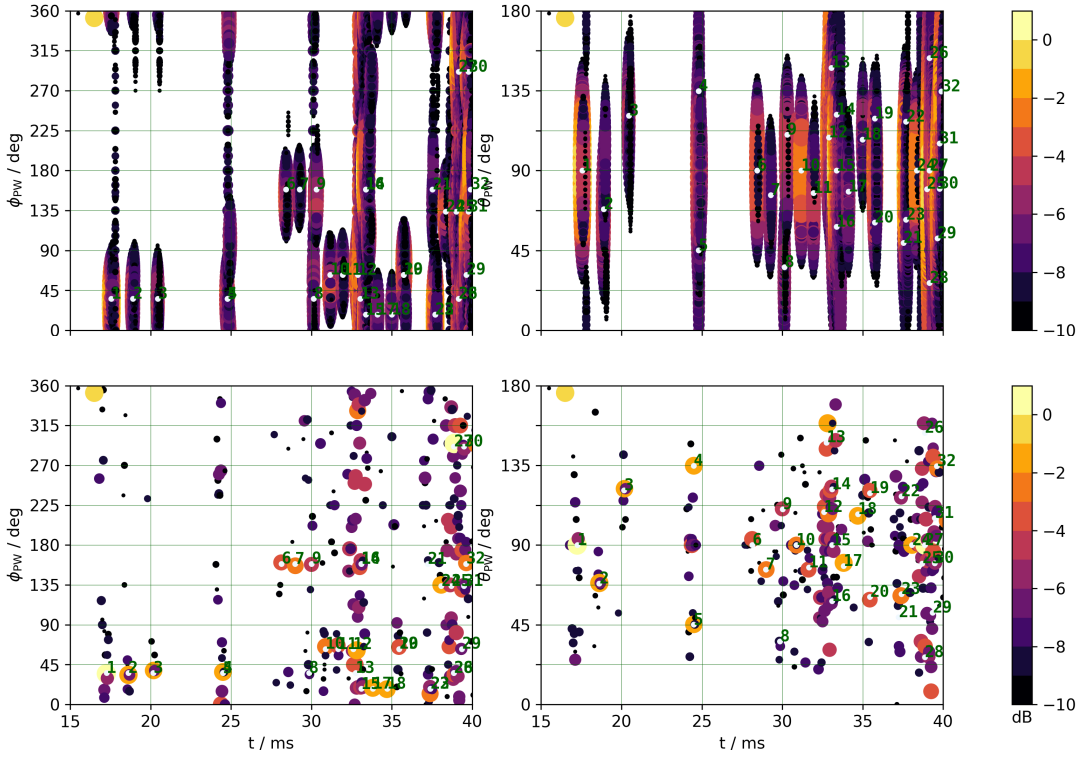

Figure 4. Early room reflections' DOA over time. Left: azimuth, right: colatitude, top: PWD $|\mathbf{p}_{\phi_m \times \theta_m \times t}|$ in dB, bottom: CS-OMP $|\mathbf{x}_{\phi_m \times \theta_m \times t}|$ in dB. Rigid microphone array, spherical Lebedev grid with 14 receivers, 4th order image source model, all walls brick (absorption coefficient 0.15), no additive noise.

for DOA estimation with proper parametrization, which is already known in literature. However, the implemented compressive sensing approach not necessarily yields outperforming results compared to traditional plane wave decomposition. Especially, when aiming at the exact count and robust DOA of early room reflections, compressive sensing might be not the optimal solution. This is not a flaw of the technique itself, but rather due violation of prerequisites, such as low mutual coherence of the sensing matrix and signal sparsity in the problem under discussion. For both, degrees of freedom in changing parameters is rather low.

REFERENCES

[1] Jarrett, D.P.; Habets, E.A.P.; Naylor, P.A. Theory and Applications of Spherical Microphone Array Processing. Springer, (CH), 2017.

[2] Rafaely, B. Fundamentals of Spherical Array Processing. Springer, Berlin (GER), 2015.

[3] Elad, M. Sparse and Redundant Representations. Springer, NY (USA), 2010.

[4] Gerstoft, P.; Mecklenbräuker, C.F.; Seong, W.; Bianco, M. Introduction to compressive sensing in acoustics. J. Acoust. Soc. Am., Vol 143 (6), 2018, pp 3731-3736.

[5] Xenaki, A.; Gerstoft, P.; Mosegard, K. Compressive beamforming. J. Acoust. Soc. Am., Vol 136 (1), 2014, pp 260-271.

[6] Xenaki, A.; Boldt, J.B.; Christensen, M. G. Sound source localization and speech enhancement with sparse Bayesian learning beamforming. J. Acoust. Soc. Am., Vol 143 (6), 2018, pp 3912-3921.

[7] Mignot, R.; Daudet, L.; Ollivier, F. Room Reverberation Reconstruction: Interpolation of the Early Part Using Compressed Sensing. IEEE Audio, Speech, Language Process., Vol 21 (11), 2013, pp 2301-2312.

[8] Verburg, S.A.; Fernandez-Grande, E. Reconstruction of the sound field in a room using compressive sensing. J. Acoust. Soc. Am., Vol 143 (6), 2018, pp 3770-3779.

[9] Spors, S.; Rettberg, T. On the Estimation of Acoustic Reflection Coefficients from In-Situ Measurements using a Spherical Microphone Array. Proc. of the 44th Deutsche Jahrestagung für Akustik (DAGA), Munich (GER), March 2018, pp 1326-1329.

[10] Lindau, A.; Kosanke, L.; Weinzierl, S. Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses. J. Audio Eng. Soc., Vol 60 (11), 2012, pp 889-898.

[11] Hahne, W; Erbes, V.; Spors, S. On the Perceptually Acceptable Noise Level in Binaural Room Impulse Responses. Proc. of the 45th Deutsche Jahrestagung für Akustik (DAGA), Rostock (GER), March 2019, pp 615-618.

5 ACKNOWLEDGEMENT

We highly appreciate and support reproducible and open science.

`https://github.com/spatialaudio/doa-early-room-reflections-pwd-vs-omp` contains a Jupyter notebook related to the present paper and the shown results.

We would like to thank Till Rettberg for his contributions to the `https://github.com/spatialaudio/sfa-numpy` toolbox.

```
#   phi/deg theta/deg  r/m    t/ms    #image source order
1  , 36 ,   90 ,    6.022 ,   17.556, 1*
2  , 36 ,   68 ,    6.482 ,   18.899, 1
3  , 36 ,   121 ,   7.016 ,   20.454, 1
4  , 36 ,   135 ,   8.501 ,   24.783, 2
5  , 36 ,   45 ,    8.501 ,   24.783, 2
6  , 159 ,  90 ,    9.75 ,    28.425, 1*
7  , 159 ,  76 ,    10.041 ,  29.274, 2
8  , 36 ,   36 ,    10.335 ,  30.132, 3
9  , 159 ,  110 ,   10.393 ,  30.301, 2
10 , 63 ,   90 ,    10.689 ,  31.164, 1*
11 , 63 ,   77 ,    10.955 ,  31.94,  2
12 , 63 ,   109 ,   11.279 ,  32.884, 2
13 , 36 ,   148 ,   11.332 ,  33.039, 3
14 , 159 ,  122 ,   11.448 ,  33.377, 3
15 , 18 ,   90 ,    11.448 ,  33.377, 1*
16 , 159 ,  58 ,    11.448 ,  33.377, 3
17 , 18 ,   78 ,    11.697 ,  34.102, 2
18 , 18 ,   107 ,   12.001 ,  34.988, 2
19 , 63 ,   119 ,   12.258 ,  35.738, 3
20 , 63 ,   61 ,    12.258 ,  35.738, 3
21 , 159 ,  49 ,    12.869 ,  37.52,  4
22 , 18 ,   118 ,   12.925 ,  37.683, 3
23 , 18 ,   62 ,    12.925 ,  37.683, 3
24 , 134 ,  90 ,    13.155 ,  38.353, 2
25 , 134 ,  80 ,    13.372 ,  38.986, 3
26 , 36 ,   153 ,   13.426 ,  39.143, 4
27 , -69 ,  90 ,    13.426 ,  39.143, 1*
28 , 36 ,   27 ,    13.426 ,  39.143, 4
29 , 63 ,   52 ,    13.595 ,  39.635, 4
30 , -69 ,  80 ,    13.639 ,  39.764, 2
31 , 134 ,  105 ,   13.639 ,  39.764, 3
32 , 159 ,  135 ,   13.683 ,  39.892, 4
```

Figure 5. All DOA of the room under discussion for $t < 40\,\text{ms}$ calculated by 4th order image source model (direct sound #1 and reflections #2-#32). * indicates horizontal 1st order mirror image sources for scenario 1. Angles rounded to integer for clarity.