Convolutional sparse formulation of the sound field in an enclosure

Hahmann, Manuel; Verburg Riezu, Samuel Arturo; Fernandez Grande, Efren

Published in:
Proceedings of inter-noise 2020

Publication date:
2020

Document Version
Peer reviewed version

Seoul inter-noise 2020 23-26 AUGUST

Convolutional sparse formulation of the sound field in an enclosure

Manuel Hahmann[1]
Technical University of Denmark,
Ørsteds Plads 352, DK-2800 Lyngby

Samuel Arturo Verburg-Riezu
Technical University of Denmark,
Ørsteds Plads 352, DK-2800 Lyngby

Efren Fernandez-Grande
Technical University of Denmark,
Ørsteds Plads 352, DK-2800 Lyngby

ABSTRACT

Sound field analysis methods often make use of elementary wave expansions to describe the sound field in a room. A widespread rationale is to express the observed sound field as a superposition of plane, spherical or other wave bases. This is frequently a good approach when considering small apertures and when the spatial variations of the sound field are similar to the chosen wave base. Yet, when considering distributed measurements over a large aperture, it becomes challenging to find analytical models that match the measured data. In practice, global wave bases can hardly account for complex sound fields that include high modal density, diffraction or scattering. This study examines methods to model the sound field in an enclosure from distributed experimental data. A methodology is proposed to extract the spatio-temporal properties of the sound field in a convolutionally sparse framework. To reduce model mismatch, it expresses the sound field as a set of local spatial patches that conform to the global data set. The technique is also suitable for approaching the problem as a spectro-spatial one.

1. INTRODUCTION

Distributed microphone array measurements enable analysis of spatio-temporal properties of sound fields. The data measured with the array is often mapped onto a linear superposition of basis functions, which are used to analyze or reconstruct the sound field. Common types of such basis functions include plane [1, 2] or spherical waves [3, 4] that originate from simple and convenient analytic expressions. Typically, one would choose a set of basis functions representative of the sound field at hand. For example for spatial sound field analysis, basis functions similar to the expected waves encountered are particularly suitable [5, 6]. In enclosures at mid-high frequencies, the complexity of the sound field increases drastically, and the use of simple elementary wave expansions spanning the whole measurement domain can be limiting [1, 2].

Theoretically, any solution to the Helmholtz equation can be represented by an infinite sum of elementary basis functions [5, 7, 8]. In sound field reconstruction, this solution is unknown and the set

[1] manha@dtu.dk

of basis functions is typically finite. For viable reconstructions, a large enough number of functions is required to represent the sound field under examination. This limits practical applications of sound field analysis with global basis functions to small domains and fields of reduced complexity. Sound fields over a large domain or at high frequencies in enclosures, possibly including local phenomena such as sources, can hardly be approximated by elementary global representations [1, 2, 9]. Analyses of such fields are therefore either locally restricted or employ a tailored function base [9, 10]. Yet, it is possible to conduct a global analysis of large, complex sound fields by partitioning the field into smaller sub-domains of lower complexity.

So called patch-based methods decompose the sound field over the global domain spatially into smaller sub-domains. Representations within these sub-domains, so called *patches*, are treated independently. All local patches then compose in sum the global field in an overlap-add fashion [11]. Independent representations in local sub-domains however neglect self-similarity between overlapping and neighboring regions. Latent structures reaching beyond a patch size might be lost. When considering a sound field, it is reasonable to assume some degree of similarity and continuity between neighboring and overlapping regions. A convolutional formulation of a sound field closes this "local-global gap" by conducting a joint analysis over all local partitions of the global domain. The local representation of a specific patch is then not only determined from measurements contained in the patch itself, but also from representations, and hence measurements, in all overlapping patches. The global representation is thus a sum of local filters convolved with their respective global coefficient map.

In the present study, we propose a convolutional formulation to analyze the reverberant sound field in a conventional room. We test the convolutional formulation reconstructing the sound field from a subset of the full experimental reference. We compare with reconstructions using conventional global plane wave expansion and locally independent patch-based method with plane waves.

2. THEORY

Sound field reconstruction commonly relies on representing the sound field as a linear combination of basis functions that extend globally over the reconstruction domain

$$\mathbf{p} = \mathbf{H}\mathbf{x}\,, \tag{1}$$

where $\mathbf{p} \in \mathbb{C}^n$ is the sound pressure over space at a given frequency, $\mathbf{H} \in \mathbb{C}^{n\times m}$ is the set of global basis functions and $\mathbf{x} \in \mathbb{C}^m$ are the coefficients. For a given set of coefficients, it is straight-forward to reconstruct by evaluating Equation 1 at desired positions. Finding the coefficients $\mathbf{x}$ is non-trivial as the corresponding inverse problem is often underdetermined, $n < m$, and rank-deficient. If suitable basis functions are chosen, one can assume that the coefficient vector is sparse. Promoting sparse solutions enables finding the coefficient estimate $\hat{\mathbf{x}}$ using far fewer measurements than required for full characterization by sampling theory [12]:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \; \|\mathbf{H}\mathbf{x} - \mathbf{p}\|_2^2 + \gamma \, \|\mathbf{x}\|_1 \quad . \tag{2}$$

The inverse problem Equation 2 minimizes a cost function depending on the squared sum of the residual $\mathbf{H}\mathbf{x} - \mathbf{p}$ and a ℓ_1-norm penalty on the coefficients. The penalty on the sum of absolute values in $\mathbf{x}$ effectively promotes sparse solutions, which are favourable when only very few functions in $\mathbf{H}$ suffice to represent $\mathbf{p}$. The ℓ_1 is a relaxation of the strict ℓ_0-norm sparsity penalty which would render the optimization problem NP-hard [13]. The regularization parameter γ weighs the penalty within the objective against the data fidelity term.

For complex problems, it can be difficult to find the globally suitable basis functions required to express $\mathbf{p}$ as Equation 1. Alternatively, the sound field $\mathbf{p}$ can also be understood as convolution of a

dictionary of k local filters $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k] \in \mathbb{C}^{d \times d \times k}$ and a global, spatially varying coefficient map $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k] \in \mathbb{C}$:

$$\mathbf{p} = \sum_k \mathbf{d}_k * \mathbf{x}_k \quad . \tag{3}$$

In other words, Equation 3 expresses the sound field as a sum of local representations, removing the need for basis functions that are globally suitable over the reconstruction domain [11,14]. This allows for representations of almost arbitrarily large domains and high complexity. Using the convolutional decomposition of $\mathbf{p}$, the inverse problem finds the estimate of the coefficient map $\hat{\mathbf{X}}$ as

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{x}} \frac{1}{2} \left\| \sum_k \mathbf{d}_k * \mathbf{x}_k - \mathbf{p} \right\|_2^2 + \gamma \sum_k \|\mathbf{x}_k\|_1 \ . \tag{4}$$

Parallel to Equation 2, we find the optimal coefficients by minimizing the sum of squared residuals and the sparsity promoting ℓ_1 penalty weighted by regularization parameter γ.

Sound field reconstructions benefit from certain adaptations of the standard convolutional sparse coding scheme. As measurements in a sound field reconstruction problem can be sparse and randomly distributed, only very few measurements might be available to estimate coefficients at one position. The convolution forms the sound field over the complete domain, $\hat{\mathbf{p}} \in \mathbb{C}^{n_p} = \sum_k \mathbf{d}_k * \hat{\mathbf{x}}_k$. A binary spatial mask $\mathbf{m} \in \{0, 1\}^{n_p}$ allows computing the residual exclusively at the measurement positions. This technique is also referred to as mask decoupling [15, 16]. In effect, $\mathbf{m}$ assumes 1 at indices where measurements are available and 0 elsewhere. Further, it seems fair to assume that coefficients at two neighboring positions only vary slightly, e.g. that sound fields only vary slowly within short distances. This observation can be reflected in the objective function by penalize the norm of the spatial coefficient gradient $\|\nabla \mathbf{x}\|_2^2$.

The coefficient map of the convolutional representation can be found by minimizing the cost function

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{x}} \frac{1}{2} \left\| \mathrm{diag}(\mathbf{m}) \left(\sum_k \mathbf{d}_k * \mathbf{x}_k - \mathbf{p} \right) \right\|_2^2 + \gamma \sum_k \|\mathbf{x}_k\|_1 + \frac{\mu}{2} \sum_k \|\nabla \mathbf{x}_k\|_2^2 \quad , \tag{5}$$

where the mask $\mathbf{m}$ is applied to evaluate the residual only at the measurement positions. It is required as the convolution is continuous throughout the complete domain. The mask is implicit in problems of the algebraic form as in Equation 2, where only measurement positions contribute to the residual. Penalties on sparsity and spatial gradient of the coefficient map are weighted by regularization parameters γ and μ respectively.

3. SETUP

This study compares global, patch-based and convolutional representations by reconstructing a sound field in a classroom over a $1.7 \times 1.7\mathrm{m}^2$ 2D domain. The experimental reference was acquired with 2.5 cm resolution over the complete domain, in total $69 \times 69 = 4761$ microphone positions. From a sparse set of pseudo-random positions, we reconstruct over the complete domain and assess the results against the measured reference. Figure 1 shows a sketch with dimensions and a photograph of the room during the measurement campaign. The classroom has ceiling base absorbers and has a fairly constant reverberation time around 500 ms through all 1/3rd octave bands.

All sound field formulations in this comparison, global, patch-based and convolutional, employ plane waves with directions sampled in three dimensions from a Fibonacci sphere. In the global representations, a dictionary of 1000 waves is available to find the best reconstruction via Equation 1. Both patch-based and convolutional representations use the same dictionary of 100 plane wave functions with a side length of two wavelengths. 40 of such local plane waves are shown in Figure 2,

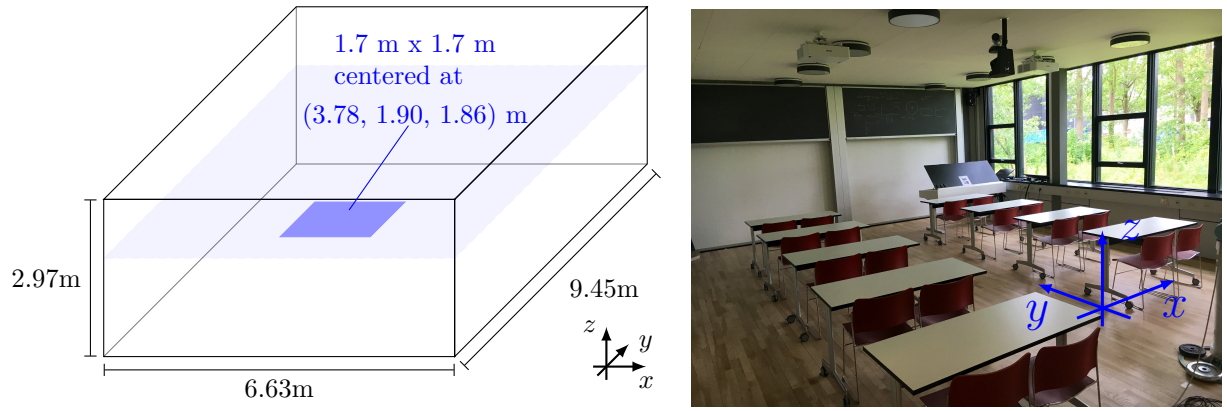


Figure 1: Left: sketch of the classroom with square measurement aperture in blue. Right: photograph of the classroom during measurement campaign.

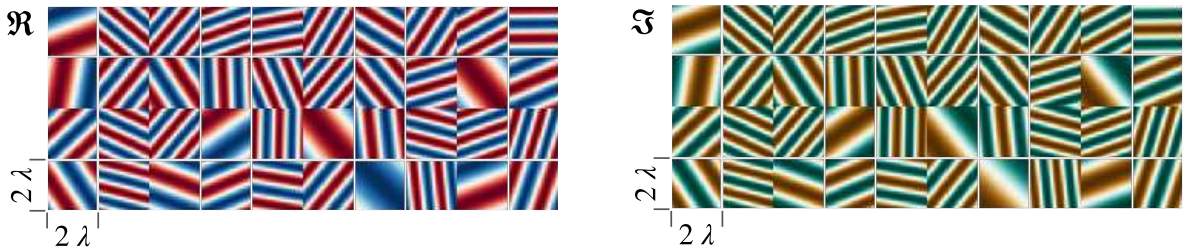


Figure 2: Local plane wave filters used in this study, real and imaginary part (left and right). Showing 40 out of 100 filters with directions distributed on a 3D Fibonacci sphere and of side length two wavelengths.

in real and imaginary part. Note how spatial frequencies of the 2D filters can be below the temporal frequency, depending on the wave propagation component in z-direction.

On a high level assessment, we compare the reconstruction against the measured reference in terms of modal assurance criterion

$$\mathrm{MAC} = |\mathbf{p}_{\mathrm{rec}}\mathbf{p}_{\mathrm{ref}}^{\mathrm{H}}|^2 \, (\mathbf{p}_{\mathrm{rec}}\mathbf{p}_{\mathrm{rec}}^{\mathrm{H}})^{-1} \, (\mathbf{p}_{\mathrm{ref}}\mathbf{p}_{\mathrm{ref}}^{\mathrm{H}})^{-1} \quad , \tag{6}$$

which expresses spatial similarity between 0 (no similarity) and 1 (identical) and normalized mean square error,

$$\mathrm{NMSE} = 10 \log_{10} \left\langle |\mathbf{p}_{\mathrm{rec}} - \mathbf{p}_{\mathrm{ref}}|^2 \, |\mathbf{p}_{\mathrm{ref}}|^{-2} \right\rangle \quad , \tag{7}$$

where $\langle \cdot \rangle$ denotes the spatial average.

Efficient solving of Equation 3 in the frequency domain using the convolution theorem imposes circular boundary conditions on the sound field $\mathbf{p}$. Spatially periodic sound fields are rare and convolutional representations of general sound fields would be compromised by artifacts around boundary. A treatment is, just as in conventional convolution, to pad the domain boundary by filter size $d - 1$. For a sound field on a two dimensional, square regular grid with $\sqrt{n} \times \sqrt{n}$ measurements, padding with $d - 1$ expands to $n_p = n + 4\sqrt{n}\,(d - 1) + 4(d^2 - 2d + 1)$ positions. To reduce the discontinuity around the boundary, we pad with reflected values such that two elements on either side of the boundary position have equal value. Correspondingly, zero-padding of the same extent has to be applied to the measurement mask $\mathbf{m}$.

The optimization problem of the convolutional formulation is rewritten as an ADMM problem [17] and solved via the `sporco.admm` software module [18]. Real and imaginary parts were treated separately. Regularization parameters were all chosen via a grid search for the respective test condition and the setting yielding the highest Equation 6 was selected. The microphone positions are

pseudo-random, i.e. uniform randomly distributed with a minimum distance to the nearest neighbour. Several iterations were run for each condition to monitor sensitivity to the precise sampling positions.

4. RESULTS

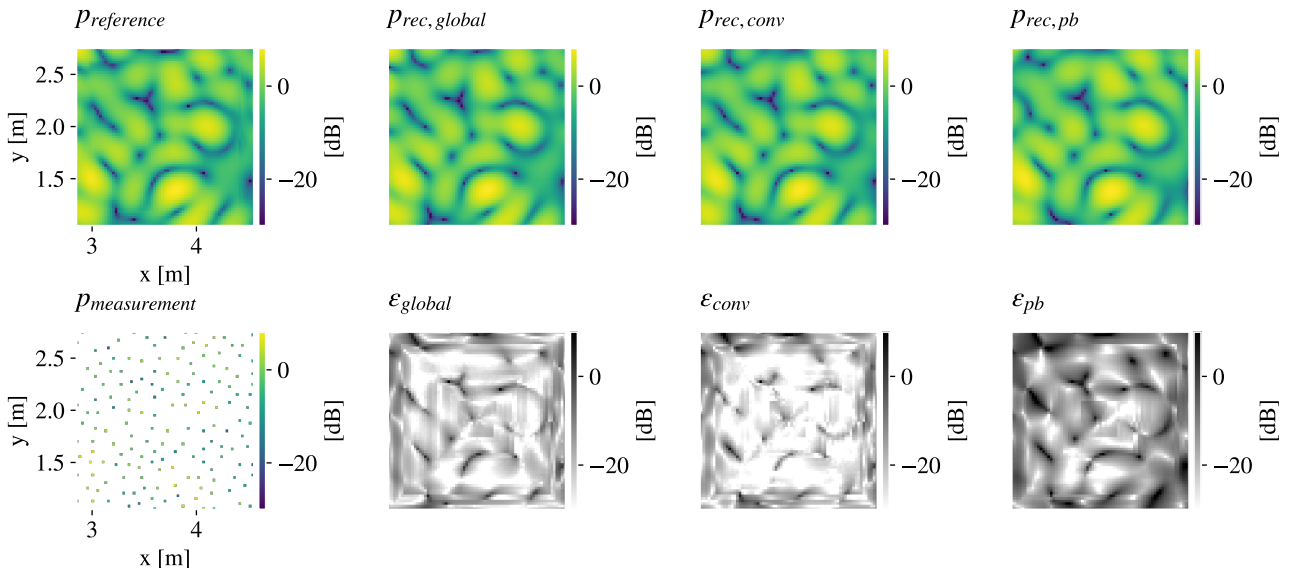


Figure 3: Reconstruction from 160 measurements at 600 Hz. From top to bottom, left column: reference and measurements for reconstruction; center left: reconstruction with 1000 global plane waves (MAC Equation 6 0.98, NMSE Equation 7 -11.2 dB); center right: reconstruction with 100 local plane wave filters, convolutional formulation (MAC 0.99, NMSE -11.0 dB); right column: reconstructions with 100 local plane waves, patch-based formulation (MAC 0.95, NMSE -4.7 dB). Pressure fields in top row and left column in dB relative to the spatial mean squared pressure. Error maps in the right column in dB relative to the reference. All subplots have the same x- and y-axis, shown in the bottom left subplot only for better overview.

Figure 3 shows reconstructions of the sound field in the classroom at 600 Hz (top left) from 160 measurements (bottom left). All figures share the same x- and y-axis, and all pressure and error level plots use aligned color scales. For each method, the reconstructed sound field (top) and error map (bottom) are shown. The global plane wave representation (center left column) achieves very good reconstruction quality (MAC 0.98, NMSE -11.2 dB). Its error map reveals a square pattern of artifacts. The convolutional model (center right column) yields a high quality reconstruction as well (MAC 0.99, NMSE -11.0 dB). The error map shows very few artifacts and low errors, especially in the center of the domain. The patch-based reconstruction (right column) has an around 5 dB higher error than both other methods (MAC 0.95, NMSE -4.7 dB). Errors are located mainly in regions around pressure level minima. With only 100 local plane wave filters, the convolutional reconstruction has similar quality than 1000 global plane waves. Sufficiently far away from the boundaries, it is flexible enough to represent and reconstruct even local phenomena with high precision. There is a clear benefit from the global joint analysis by the convolutional form compared to independent local representations in the patch-based method.

Figure 4 shows reconstructions of the sound field in the classroom at 1250 Hz (top left) from 320 measurements (bottom left). All figures share the same x- and y-axis, and all pressure and error level plots use aligned color scales. As above, reconstructed sound fields and normalized error maps are shown for the three representations: global (center left), convolutional (center right) and patch-based method (right column). All reconstructions seem plausible, but show higher errors compared to the 600 Hz case, especially around boundary and pressure level minima regions. The convolutional

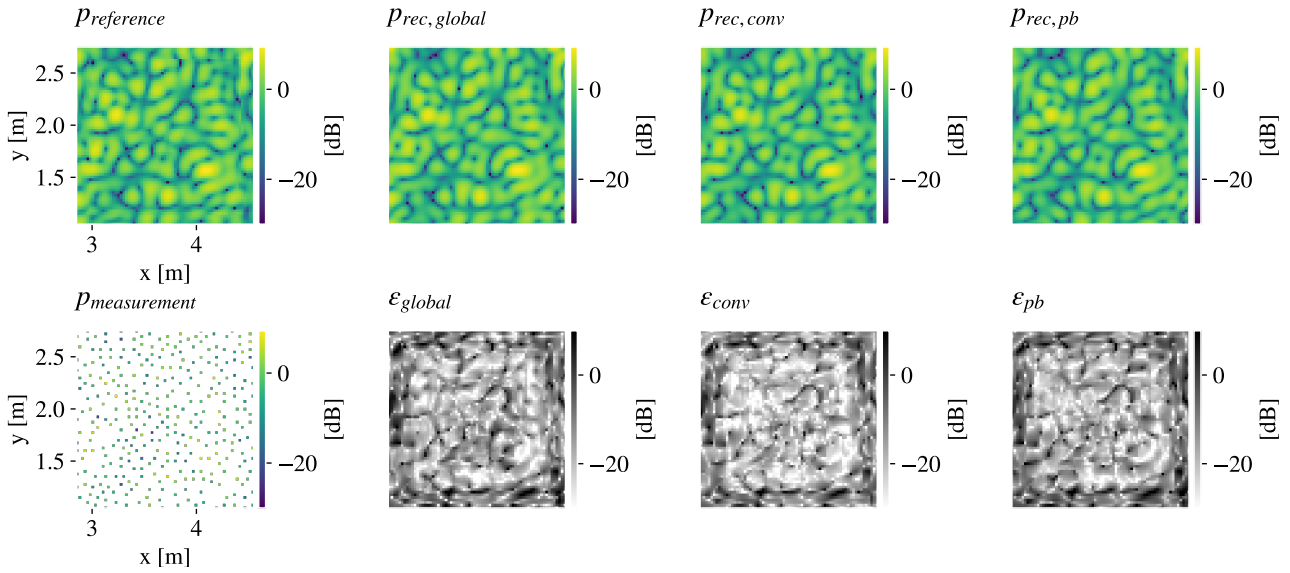


Figure 4: Reconstruction from 320 measurements at 1250 Hz. From top to bottom, left column: reference and measurements for reconstruction; center left: reconstruction with 1000 global plane waves (MAC Equation 6 0.92, NMSE Equation 7 -0.5 dB); center right: reconstruction with 100 local plane wave filters, convolutional (MAC 0.95, NMSE -2.6 dB); right column: reconstructions with 100 local plane waves, patch-based formulation (MAC 0.92, NMSE -1.52 dB). Pressure fields in top row and left column in dB relative to the spatial mean squared pressure. Error maps in the right column in dB relative to the reference. All subplots have the same x- and y-axis, shown in the bottom left subplot.

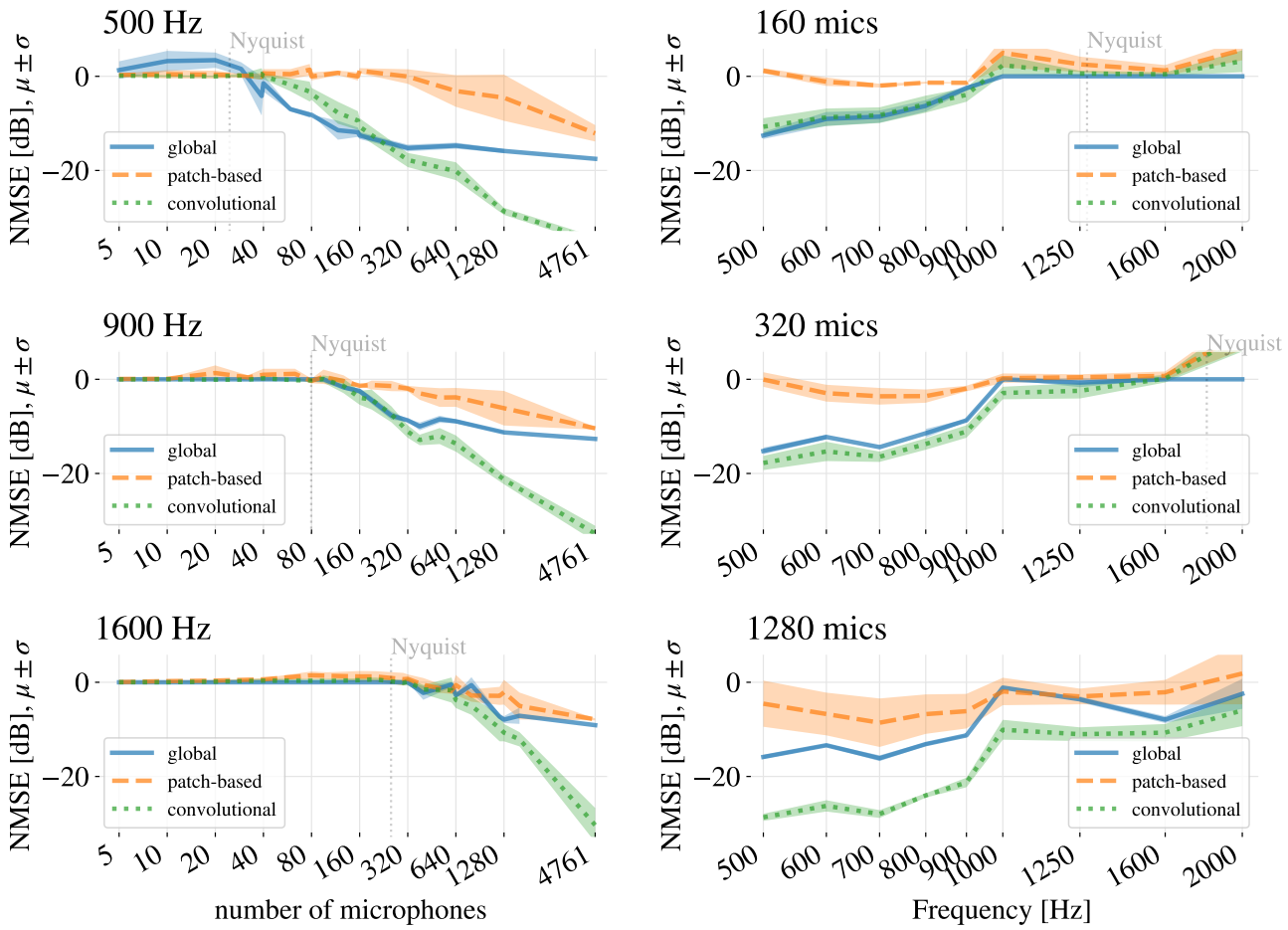
reconstruction has slightly lower error (NMSE -2.7 dB) than the patch-based (-1.5 dB) and global method (-0.5 dB). The 1000 global plane waves do not seem able to represent local minima quite as well as the convolutional model. The convolutional reconstruction reaches high precision with sufficient measurements and distance to the boundary.

In a more comprehensive study, we examine the reconstructions for a varying number of microphones and as a function of frequency. Figure 5 shows the reconstruction quality in terms of NMSE Equation 7 depending on number of microphones at fixed frequencies (Figure 5a) and conversely over a frequency range between 500 Hz and 2 kHz for fixed number of microphones (Figure 5b). Whenever it falls within the abscissa range, the approximate Nyquist sampling limit is indicated, i.e. on average one microphone per half wavelength and dimension, $(\lambda/2)^{\text{dim}}$.

Some general trends for all methods can be observed in Figure 5a. Firstly and straightforwardly, viable reconstructions need a sufficient number of measurements and the more information is available, the better the reconstruction. Secondly, with increasing frequency, more measurements are required, approximately shifting gradually with the Nyquist rate.

Most notably, the patch-based method underperforms throughout the frequency range and requires many measurements to yield viable reconstructions. In contrast to the independent local representation, the joint analysis by the convolutional formulation enables reconstruction of fine spatial details and with low error. However, it does require sufficiently many measurements to benefit from the flexibility of the convolutional form. Also global plane waves achieve accurate reconstructions at higher frequencies, but seem to meet a limit below which even more measurements do not yield any more improvement.

At higher reconstruction qualities, the NMSE Equation 7 in Figure 5b is more descriptive than the MAC Equation 6. With regards to the focus and clarity of this paper, corresponding illustrations of MAC are therefore omitted. The main observation is there, that when the number of microphone positions is close to or below the Nyquist rate, the global formulation is the most suitable one and



a) *b)*

Figure 5: Normalised mean square error Equation 7 of reconstructions, using —global, - - -patch-based or ·····convolutional formulations of the sound field. Shown is one standard deviation σ around the repetition mean μ over pseudo-random measurement locations. All subplots share the same y- and all columns the same x-axes: a) NMSE across number of microphones at specific frequencies and b) NMSE across frequencies at fixed numbers of microphones.

provides better reconstructions than both other methods.

From above experiments we see that expressing a sound field in terms of a convolution of a coefficient map with only few local filters yields flexible and precise representations when provided with sufficient measurements. While the studied cases were fairly homogeneously sampled and without very prominent local phenomena, convolutional coding seems also to have potential in analysis of sound fields with heterogeneous spatial properties. The square shaped artifacts observed in reconstructions in both Figure 3 and 4 are in fact present in the experimental reference data. They seem to originate from variations in ambient conditions during the approximately 40 hour long acquisition time. The convolutional formulation even partly reconstructs those local artifacts, indicating its flexibility.

It is to note that the best reconstructions with the convolutional formulation were achieved for low γ and high μ values in Equation 5. This corresponds to promoting smoothness in the coefficients rather than sparsity. One could argue that the choice of a dictionary with only 100 plane waves already is inherently sparse design for a sound field of such complexity. Under certain circumstances though, enforcing a sparse representation might be beneficial and improve reconstructions where only few samples are available. That could include filters from other suitable basis functions, varying their number and extent. In fact, learning the most optimal set of filters is a main motivation behind both

patch-based methods and convolutional sparse coding [19, 20].

Compared to typical convolutional sparse coding problems in image processing like superresolution inpainting and denoising, only very few measurements are available in sound field reconstruction tasks. It is therefore crucial that the local filters have spatial properties common to sound fields. Choosing plane wave filters corresponds to assuming an upper limit on the spatial frequency that can physically occur in the sound field. In addition, regularization of the coefficient gradient imposes similarity between neighboring patches in the sound field, which is appropriate from a physical standpoint. Additionally, using filters originating from the analytical expression facilitate straight-forward reconstruction of the particle velocity or sound intensity fields and other analyses conducted with traditional plane wave representations.

5. CONCLUSIONS

In this study we have proposed and examined the use of convolutional formulations of a sound field in space domain. The examined method consists in dividing a given reconstruction domain into smaller patches and then expanding the measured data onto a convolutional sparse model, that accounts for the local and global structures present in the data. The method has been tested by reconstructing experimentally the sound field in a conventional room. The results indicate that the methodology achieves a highly flexible and precise representation of an arbitrary sound field, and is especially beneficial whenever common basis functions cannot accurately map the observed data. The convolutional form examined captures fine details of local sound field phenomena and expresses them in terms of only a few local functions. The experimental results also show that convolutional formulations are very accurate interpolators whenever a sufficient amount of data are available.

6. ACKNOWLEDGEMENTS

The authors would like to thank Villum Fonden for financial support through grant number 19179 for the project ‘Large-scale Acoustic Holography’.

References

[1] S. A. Verburg and E. Fernandez-Grande, “Reconstruction of the sound field in a room using compressive sensing,” *Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3770–3779, 2018.

[2] R. Mignot, G. Chardon, and L. Daudet, “Low frequency interpolation of room impulse responses using compressed sensing,” *Ieee Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 205–216, 2014.

[3] R. Mignot, L. Daudet, and F. Ollivier, “Room reverberation reconstruction: Interpolation of the early part using compressed sensing,” *Ieee Transactions on Audio, Speech and Language Processing*, vol. 21, no. 11, pp. 6562745, 2301–2312, 2013.

[4] E. Fernandez-Grande, “Sound field reconstruction using a spherical microphone array,” *Journal of the Acoustical Society of America*, vol. 139, no. 3, pp. 1168–1178, 2016.

[5] E. G. Williams, *Fourier acoustics : sound radiation and nearfield acoustical holography*. Academic Press, 1999.

[6] E. Fernandez-Grande, A. Xenaki, and P. Gerstoft, “A sparse equivalent source method for near-field acoustic holography,” *Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 532–542, 2017.

[7] A. Moiola, R. Hiptmair, and I. Perugia, "Plane wave approximation of homogeneous helmholtz solutions," *Zeitschrift Fur Angewandte Mathematik Und Physik*, vol. 62, no. 5, pp. 809–837, 2011.

[8] G. Koopmann, L. Song, and J. Fahnline, "A method for computing acoustic fields based on the principle of wave superposition," *Journal of the Acoustical Society of America*, vol. 86, no. 6, pp. 2433–2438, 1989.

[9] T. Nowakowski, J. De Rosny, and L. Daudet, "Robust source localization from wavefield separation including prior information," *Journal of the Acoustical Society of America*, vol. 141, no. 4, pp. 2375–2386, 2017.

[10] M. Hahmann, S. A. Verburg Riezu, and E. Fernandez-Grande, "Analysis of a sound field in a room using dictionary learning," pp. 149–154, 2019. 23rd International Congress on Acoustics, 9 - 13 September 2019, Aachen.

[11] M. Elad, *Sparse and redundant representations: From theory to applications in signal and image processing*. Springer New York, 2010.

[12] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *Ieee Transactions on Signal Processing*, vol. 54, no. 10, pp. 3790–3804, 2006.

[13] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *Ieee Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[14] V. Papyan, Y. Romano, and M. Elad, "Convolutional neural networks analyzed via convolutional sparse coding," *Journal of Machine Learning Research*, vol. 18, pp. 1–52, 2017.

[15] F. Heide, W. Heidrich, and G. Wetzstein, "Fast and flexible convolutional sparse coding," *Proceedings of the Ieee Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-, pp. 7299149, 5135–5143, 2015.

[16] B. Wohlberg, "Boundary handling for convolutional sparse representations," *Proceedings - International Conference on Image Processing, Icip*, vol. 2016-, pp. 7532675, 1833–1837, 2016.

[17] M. S. C. Almeida and M. A. T. Figueiredo, "Deconvolving images with unknown boundaries using the alternating direction method of multipliers," *Ieee Transactions on Image Processing*, vol. 22, no. 8, pp. 3074–3086, 3074–3086, 2013.

[18] B. Wohlberg, "SParse Optimization Research COde (SPORCO)." Software library available from http://purl.org/brendt/software/sporco, 2016.

[19] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Ieee Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[20] V. Papyan, Y. Romano, J. Sulam, and M. Elad, "Theoretical foundations of deep learning via sparse representations a multilayer sparse model and its connection to convolutional neural networks," *Ieee Signal Processing Magazine*, vol. 35, no. 4, pp. 72–89, 2018.