ICA 2019 AACHEN

PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Analysis of a sound field in a room using dictionary learning

Manuel HAHMANN[(1)], Samuel A. VERBURG RIEZU[(2)], Efren FERNANDEZ-GRANDE[(3)]

[(1)]Acoustic Technology, DTU Elektro, Denmark, manha@dtu.dk
[(2)]Acoustic Technology, DTU Elektro, Denmark, saveri@elektro.dtu.dk
[(3)]Acoustic Technology, DTU Elektro, Denmark, efg@elektro.dtu.do

Abstract

The sound field in a room is often modeled as a superposition of elementary waves, such as plane or spherical waves. These wave expansions provide a powerful means to interpolate or extrapolate the sound field within (and outside) the measurement domain. However, projecting the sound field of a large domain in a room on a planar or spherical wave base yields a high number of very elemental components. We examine the use of dictionary learning to find a set of alternative basis functions that are suitable to represent the sound field enclosed in a room. The resulting dictionary is able to capture the dominant features of the sound field, and represent it using only a sparse set of functions, the dictionary atoms. In this study, high resolution measurements of the sound pressure in a room are simulated and used as a training set to learn a dictionary. We analyze the spatial properties of the learned dictionary, and compare it to simple elementary basis functions such as plane and spherical waves.

Keywords: Sound Field Reconstruction, Room Acoustics, Dictionary Learning

1 INTRODUCTION

Sound field reconstruction techniques often rely on the use of basis functions to represent sound fields. Normally, elementary wave functions (plane and spherical waves) are employed [1–4]. Representing a sound field using wave functions make it possible to interpolate the sound field in the measurements area, and even to extrapolate it outside. Wave functions contain physical meaning regarding the propagation of sound since they are simple solutions to the wave equation. proof that specific solutions to the homogeneous wave equation can be approximated by a sum of plane waves [5] with certain guaranties on the quality. However, representing complex sound fields across large three-dimensional domains, such as rooms, using elementary waves might be sub-optimal. The number of wave functions required to approximate a sound field increases with the square of the frequency and characteristic length of the room, e.g. more than 3000 plane waves would be required to approximate a sound field in a $3 \times 3 \times 3$ m volume up to 1 kHz [6]. The ideal modal behaviour of the sound field in rectangular enclosures at low frequencies makes it possible to assume sparsity, i.e. just a few waves (8 in the case of an oblique mode [7]) are non-zero. Several studies [8–12] make use of the sparsity assumption to alleviate sampling requirement. However, complex sound fields in real rooms are not sparse. Distributed sound sources, scattering and diffraction phenomena, and non-rectangular geometries are not well approximated using a small number of wave functions.

This work aims at presenting a dictionary learning (DL) approach to the problem, where basis functions are learned from densely sampled sound fields. This data driven method showed promising dimensionality reduction for similar reconstruction tasks [13, 14]. The so called atoms balance between incorporating inherent information from the data into the basis functions while preserving flexibility to sparsely represent complex sound fields. These basis functions are learned from a data set of local microphone arrays at a single frequency. They are subsequently used to reconstruct a sampled synthesized sound field at various frequencies and sampling grids. A plane wave expansion is used as reference model to evaluate the reconstruction quality.

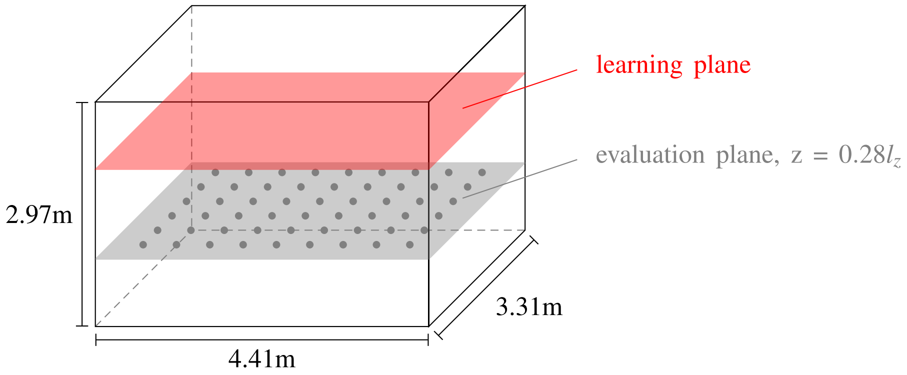

Figure 1. The cuboid room geometry in question, with a learning plane at random height (red) and an evaluation plane for method comparison (gray) at a fixed height. Dots in the evaluation plane illustrate a regular microphone grid as it is used for reconstruction input.

2 THEORY

Sparse coding methods enforce sparsity in the elements of the loading vector γ. For a projection matrix $\mathbf{A}$ and a signal x, that writes

$$\min_{\gamma} \|\gamma\|_0, \text{ s.t. } x = \mathbf{A}\gamma \quad . \tag{1}$$

This problem is known to be NP-hard, but can be approximated by greedy algorithms such as matching pursuit and thresholding [15]. Alternatively relaxation algorithms such as basis pursuit / LASSO approximate the solution by smoothing the ℓ_0 norm and solve an ℓ_1 norm optimization problem [16]:

$$\min_{\gamma} \frac{1}{2}\|x - \mathbf{A}\gamma\|_2^2 + \lambda\|\gamma\|_1 \quad , \tag{2}$$

where λ is a regularization parameter.

Dictionary learning infers the loading γ and the atoms d_i of a dictionary $\mathbf{D}$. Here, DL is performed as described in [17], alternating between a sparse coding and a learning stage. The first employs relaxation of the ℓ_0 to the ℓ_1 norm for approximation. After initializing a dictionary $\mathbf{D}$, a sparse coding step is performed on the data, optimizing with respect to γ, with $\mathbf{A} = \mathbf{D}$ in Eq. 2. Next, all atoms d_i are optimized one by one, such that the distance to the data set with nonzero γ_i coefficients is minimized. With the updated dictionary, iterations continue from the sparse coding step until defined convergence criteria or maximum iterations are reached. Subsequently, Orthogonal Matching Pursuit (OMP) is then used during reconstruction of missing data [15].

2.1 Plane wave expansion

The reference case is a plane wave basis, commonly used where maximum flexibility is needed to derive the sound field. For a given frequency, K wavenumber vectors are sampled from a sphere with radius $k = 2\pi f/c_0 = \sqrt{k_x^2 + k_y^2 + k_z^2}$. Each of these represents a global plane wave, whose complex amplitudes are found during inference. Thus any sound field could be represented if an infinite number of plane waves were available. Each mode in a reverberant, cuboid room can be represented by eight plane waves [7]. It follows that the lower bound of projection sparsity without information loss is given as eight times the modal overlap [4]. With the projection matrix denoted as $\mathbf{H}$ and the residual error e_{pwe}, it is

$$x = \mathbf{H}\gamma + e_{pwe} \quad . \tag{3}$$

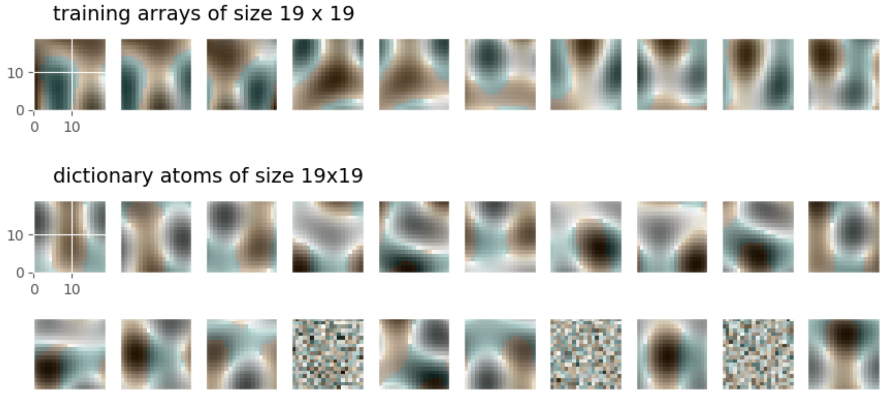

Figure 2. Top: subset of ten arrays from the learning data set. Bottom: All 20 atoms $d_{i=1...20}$ of learned complex dictionary $\mathbf{D}$, low to high magnitude from white to black and angle indicated by color overlay.

3 SIMULATION STUDY

3.1 Dictionary learning

Experiments are carried out on synthesized data, using the Greens function for a cuboid, empty and hard-walled enclosure as a reference. The dimensions of the room are $[l_x, l_y, l_z] = [4.41, 3.31, 2.97]$ m, with the virtual source placed in the corner to excite all modes. At a resolution of r, this results in a grid of $N_{mics} = (\frac{l_x}{r} + 1) \times (\frac{l_y}{r} + 1)$ microphones in any horizontal cross section. Samples are extracted such that for a array size of $P \times P$, one side length covers at least one wavelength λ. In total, $N_{samples} = (l_x/r - P + 2) \times (l_y/r - P + 2)$ sample arrays constitute the base for DL. The randomly chosen frequency of 403 Hz is about one third octave below the rooms Schroeder frequency of $f_S \approx 2000\sqrt{T_{rev}/V} \approx 500$ Hz. This yields $N_{samples} = 3479$ arrays of 19×19 microphones at 5 cm resolution to cover the whole plane. Ten samples are illustrated in the top row of Fig. 2. From those the dictionary $\mathbf{D}$ is inferred on the zero-mean aligned arrays. All 20 atoms are shown in the lower rows of Figure 2. The learned atoms seem to include the spatial features present in the training data. They contain similar spatial variations with zero crossings and maxima as expected for an aperture of size λ^2. $\mathbf{D}$ is overcomplete, containing atoms that show no spatial correlation and appear to be noise. The corresponding elements in γ are close to zero. This means that the learned dictionary is a good candidate to represent the training data sparsely. It also indicates that a target size of $N_{atoms} = 20$ is sufficient.

3.2 Sound field reconstruction

Reconstruction of the sound field in the evaluation plane is carried out using orthogonal matching pursuit (OMP) for both learned dictionary and plane wave basis, resulting in the corresponding sound pressures p_{dl} and p_{pwe}. For the global plane wave expansion, a sparse representation of $k_0 = 30$ out of $N_{pw} = 4000$ Fibonacci-spaced plane wavenumber vectors is used. The frequency dependent $N_{mics} \times N_{pw}$ transfer matrix $\mathbf{H}$ for the whole domain is constructed. OMP determines the loading vector γ of length N_{pw} and the aforementioned sparsity. $\mathbf{D}$ represents a local support of size $P \times P$, so that the reconstruction is conditioned in the same fashion as for the learning step. The sparse coding step for reconstruction of the j'th sample then involves the data x_{j,P^2}, dictionary $\mathbf{D}_{N_{atoms} \times P^2}$, yielding γ is of length N_{atoms}. The whole domain is consequently reconstructed by overlapping results of $N_{samples}$ sparse coding steps.

Figure 3 shows the reconstruction from both methods in the evaluation plane, at the same frequency as the learned dictionary. The true field in a) serves as a reference for the reconstruction from measurement grid in d). The resulting p_{dl} and plane wave reconstruction p_{pwe} are presented in b) and e) respectively. c) and f)

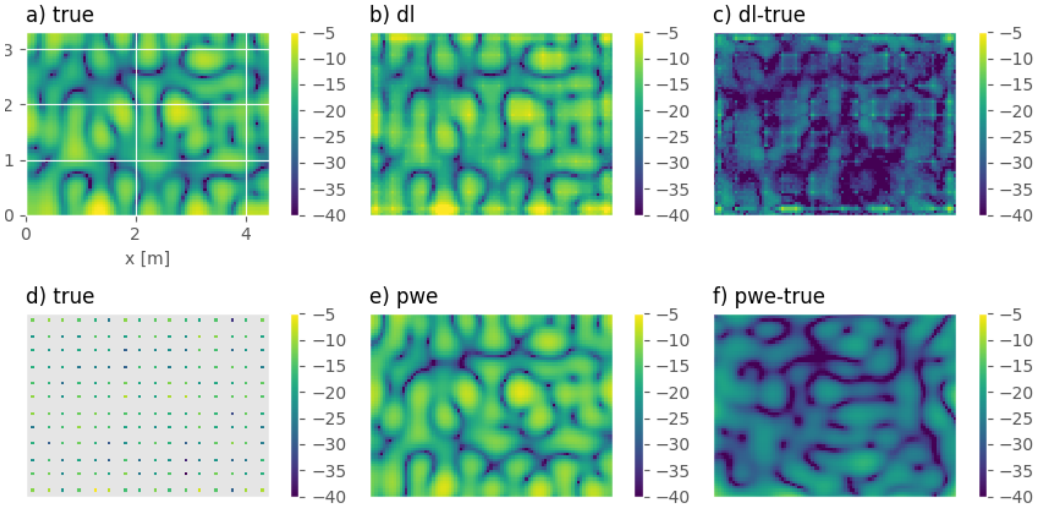

Figure 3. Reconstruction from 192 microphone measurements on a regular grid in the evaluation plane at 403 Hz. All in magnitude dB rel 1 Pa: a) true reference sound field, b) reconstruction using the learned dictionary, c) residual error between true reference and reconstruction with learned dictionary, d) regular grid sampling the true reference at 192 microphone positions for reconstruction; e) reconstruction using plane wave expansion, f) residual error between true reference and plane wave reconstruction

illustrate corresponding error magnitudes.

p_{dl} is close to the true field, both in spatial pattern and magnitude. Visible artifacts relate to the regular spacing of the grid in d) and the atom shape $\lambda \times \lambda$. This impression is confirmed when inspecting the residual error. Fig 3 c) contains visible local artifacts at measurement grid points and components of the reference field. These local variations aside, the error over space is low. p_{pwe} in Fig 3 e) shows hardly any visual differences to the true reference. The error surface in Fig 3 f) is smooth and seemingly unrelated to the true field, which in fact can be expressed a superposition of a few plane waves, given this simple geometry. Therefore it constitutes a strong reference for comparison. Its reconstruction errors originate mainly from discrete sampling of the wavenumbers and the microphone spacing.

As projection with **D** only relies on local information within a λ^2 array, it is particularly interesting to monitor global properties. The mean square relative error is shown for different numbers of microphones in Fig. 4 a). The number of microphones above which the Nyquist sampling theorem is fulfilled, i.e. a spacing of $< \lambda/2$, is marked as n. While having comparable errors for undersampled data, the dictionary projection p_{dl} converges to a higher accuracy than p_{pwe} for large N_{mics}.

Figure 4 b) and c) show the statistics of the original and reconstructed sound fields. Being slightly below the Schroeder frequency, the comparison to statistical room acoustics theory can at least give an indication of plausibility. The data follows the theoretical diffuse field pure tone model fairly well. More precisely, this is expressed by a exponentially distributed mean square pressure [**morse1968a**] in b) - note the logarithmic ordinate covering several orders of magnitudes. Likewise the sound pressure level distributions in c) follow the reference and the model well.

D here has dimensions 20×361 and 3479 coding steps are required for projection, whereas for the plane wave expansion, only a single coding step is required inverting a considerably larger matrix of $dim(\mathbf{H}) = (5963 \times 4000)$. Anecdotally, reconstruction with the dictionary took 30-100% longer than the projection on a plane wave basis for the study case. With respect to the larger objective, the question needs to be posed if the

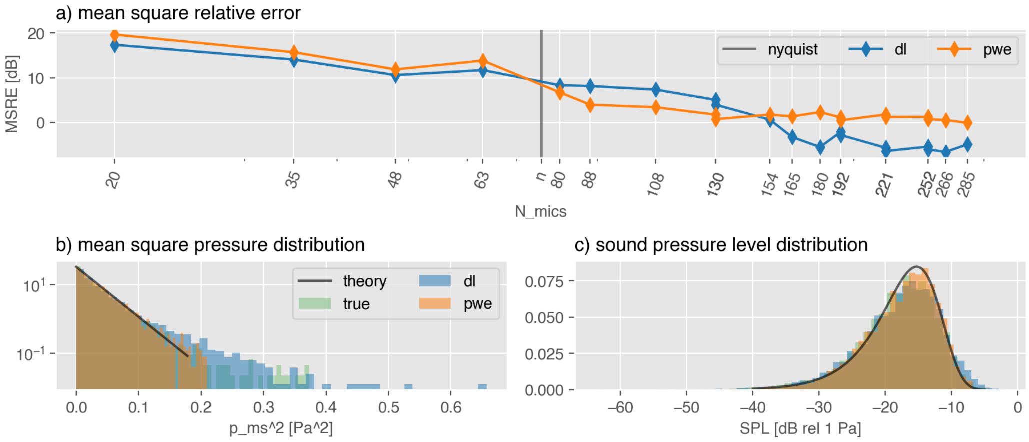

Figure 4. a): Mean square relative error depending on the number of microphones in the evaluation plane N_{mics}. b) and c): Statistics of true and reconstructed sound fields at 403 Hz, mean squared pressure and sound pressure level, along with the theoretical curve from a pure tone diffuse field model.

learned dictionary is generalizable across different rooms. Transferring the method to measured data, examining across-frequency flexibility and studying training data representativity will be part of future work.

4 CONCLUSION

A dictionary learning approach to model and reconstruct sound fields from an incomplete data set was introduced. A set of 20 basis functions, i.e. atoms were learned and sufficient to reconstruct sound fields in a cuboid room. The set formed an overcomplete dictionary, enabling a sparse representation of the training data.

Using the dictionary for sound field reconstruction, accuracies comparable to a plane wave expansion were reached. While the spatial error average was low, inherent limits in projecting a global sound field on a local dictionary became apparent in non-smooth reconstruction artifacts. The study indicates that statistical properties of the sound field are preserved.

It is the same local finiteness of the atoms that can prove beneficial for larger, more complex domains with a less homogeneous field. The online processing of atom-sized arrays distributes the computational load and possibly less demanding. That being said, projection on a plane wave basis is faster and suitable for the particular case of a cuboid room.

ACKNOWLEDGEMENTS

This work was generously supported by a research grant from the VILLUM foundation (grant nr. 19179).

REFERENCES

[1] E. G. Williams. *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography*. New York: Academic Press, 1999.

[2] B. Rafaely. “Plane-wave decomposition of the sound field on a sphere by spherical convolution”. In: *J. Acoust. Soc. Am.* 116.4 (2004), pp. 2149–2157.

[3] E. Fernandez-Grande. “Sound field reconstruction using a spherical microphone array”. In: *J. Acoust. Soc. Am.* 139.3 (2016), pp. 1168–1178.

[4] Rémi Mignot, Gilles Chardon, and Laurent Daudet. “Low frequency interpolation of room impulse responses using compressed sensing”. eng. In: *Ieee Transactions on Audio, Speech and Language Processing* 22.1 (2014), pp. 205–216. ISSN: 15587924, 15587916, 23299304, 23299290. DOI: 10.1109/TASLP.2013.2286922.

[5] A. Moiola, R. Hiptmair, and I. Perugia. “Plane wave approximation of homogeneous Helmholtz solutions”. In: *Z. Angew. Math. Phys.* 62 (2011), pp. 809–837.

[6] T. Nowakowski, J. de Rosny, and L. Daudet. “Robust source localization from wavefield separation including prior information”. In: *J. Acoust. Soc. Am.* 141.4 (2017), pp. 2375–2386.

[7] F. Jacobsen and P. M. Juhl. *Fundamentals of General Linear Acoustics*. Chap. 8: Sound in enclosures. London: Wiley, 2013.

[8] R. Mignot, L. Daudet, and F. Ollivier. “Room reverberation reconstruction: interpolation of the early part using compressed sensing”. In: *IEEE Trans. Audio, Speech, Lang. Process.* 21.11 (2013), pp. 2301–2312.

[9] R. Mignot, G. Chardon, and L. Daudet. “Low frequency interpolation of room impulse responses using compressed sensing”. In: *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 22.1 (2014), pp. 205–216.

[10] W. Jin and W. B. Kleijn. “Theory and design of multizone soundfield reproduction using sparse methods”. In: *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 23.12 (2015), pp. 2343–2355.

[11] N. Antonello et al. “Room impulse response interpolation using a sparse spatio-temporal representation of the sound field”. In: *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 25.10 (2017), pp. 1929–1941.

[12] S. A. Verburg and E. Fernandez-Grande. “Reconstruction of the sound field in a room using compressive sensing”. In: *J. Acoust. Soc. Am.* 143.6 (2010), pp. 3770–3779.

[13] Michael Bianco and Peter Gerstoft. “Dictionary learning of sound speed profiles”. eng. In: *Journal of the Acoustical Society of America* 141.3 (2017), pp. 1749–1758. ISSN: 15208524, 00014966, 01630962. DOI: 10.1121/1.4977926.

[14] Ivana Tošić and Pascal Frossard. “Dictionary learning”. eng. In: *Ieee Signal Processing Magazine* 28.2 (2011), pp. 5714407, 27–38. ISSN: 15580792, 10535888. DOI: 10.1109/MSP.2010.939537.

[15] Michael Elad. *Sparse and redundant representations: From theory to applications in signal and image processing*. eng. Springer New York, 2010, pp. 1–376. ISBN: 1441970118, 144197010x, 9781441970107, 9781441970114. DOI: 10.1007/978-1-4419-7011-4.

[16] Vardan Papyan et al. “Theoretical Foundations of Deep Learning via Sparse Representations A multilayer sparse model and its connection to convolutional neural networks”. eng. In: *Ieee Signal Processing Magazine* 35.4 (2018), pp. 72–89. ISSN: 15580792, 10535888. DOI: 10.1109/MSP.2018.2820224.

[17] Julien Mairal et al. “Online dictionary learning for sparse coding”. und. In: *Proceedings of the 26th International Conference on Machine Learning, Icml 2009* (2009), pp. 689–696.