Audio Engineering Society

Convention Paper 10402

Presented at the 149th Convention
Online, 2020 October 27-30

Sparse Audio Inpainting: A Dictionary Learning Technique to Improve Its Performance

Georg Tauböck[1], Shristi Rajbamshi[1], and Peter Balazs[1]

[1]*Acoustics Research Institute, Austrian Academy of Sciences, Austria*

Correspondence should be addressed to Georg Tauböck (georg.tauboeck@oeaw.ac.at)

ABSTRACT

The objective of audio inpainting is to fill a gap in a signal, either to be meaningful or even to reconstruct the original signal. We propose a novel approach applying sparse modeling in the time-frequency (TF) domain. In particular, we develop a dictionary learning technique which deforms a given Gabor frame such that the sparsity of the analysis coefficients of the resulting frame is maximized. A suitable modification of the SParse Audio Inpainter (SPAIN) allows to exploit the obtained sparsity gain and, hence, to benefit from the learned dictionary. Our experiments demonstrate that our methods outperforms several state-of-the-art audio inpainting techniques in terms of signal-to-noise ratio (SNR) and objective difference grade (ODG).

1 Introduction

Audio inpainting [1] denotes a signal processing technique to restore gaps, i.e., missing consecutive samples, in an audio signal, while still keeping perceptible audio artifacts as small as possible. The gaps to be filled may be classified as *short*, *medium* and *long*, depending on their duration. Typically, gaps of $\leq$ 10ms are considered as short, gaps of 10ms – 100ms are considered as medium, and gaps of $\geq$ 100ms are considered as long.

Over the last decade, audio inpainting has attracted a lot of commercial and research interest as it has many important applications. Examples include compensation of audio packet losses in communication networks [2], reconstruction of audio samples caused due to scratches in CDs or old recordings [3], and many more.

The most common approach to tackle the audio inpainting problem is to exploit prior information about the signal, i.e., to utilize reliable parts of the signal to inpaint the gaps. The reliable parts could be either exploited in the original (time) domain or with respect to some redundant representation. In the time domain setting, the methods proposed by Janssen et al [4], Oudre [5], and Etter [6] are among the first. They are based on autoregressive (AR) signal modeling; the missing samples are filled by linear prediction using autoregressive coefficients that are learned from the neighborhood of the gap. Due to their excellent performance they are still state-of-the-art.

Since the advent of sparse signal representations and compressive sensing (CS) [7], several other audio inpainting techniques that exploit sparsity have been proposed [1, 8, 9], most notably [1], which introduced the term "audio inpainting" motivated by analogous image processing tasks. These methods formulate the inpainting problem as an optimization task while taking into account that many real-world audio signals have an (approximately) sparse representation with respect to a suitable dictionary.

Motivation and Contributions. This contribution is based on the recently introduced inpainting algorithm *SParse Audio Inpainter (SPAIN)* [8]. SPAIN is an adaption of the so-called SParse Audio DEclipper (SPADE) algorithm [10] to the inpainting problem and leverages sparsity with respect to (Gabor) frames [11, 12]. Our main idea is to deform the underlying frame in order to learn a representation with increased sparsity from reliable signal parts around the gap. Furthermore, we propose a modification of the original SPAIN algorithm: instead of processing the signal segment-wise, cf. [8], we apply the algorithm to the signal parts in the neighborhood of the gap as a whole. Additionally, we replace the involved ℓ_0-norm[1] by an $\ell_{0,\infty}$-norm[1], where the supremum is taken over time. By means of the learned sparsity enhanced dictionary our modified SPAIN algorithm exhibit a significantly improved reconstruction performance.

Notation. Roman letters $\mathrm{A},\mathrm{B},\ldots$, $\mathrm{a},\mathrm{b},\ldots$, and $a,b,\ldots$ designate matrices, vectors, and scalars, respectively. $\lfloor a \rfloor$ denotes the largest integer $\leq a$. With $[\cdot]_N = [\cdot \bmod N]$ we abbreviate the modulo-N operation due to circular indexing. The ith component of the vector u is u_{i-1}; the element in ith row and jth column of the matrix A is $\mathrm{A}_{i-1,j-1}$. The superscripts $^{\mathsf{T}}$ and $^{\mathsf{H}}$ stand for transposition and Hermitian transposition, respectively. The $N \times N$ identity matrix is denoted by I_N; the $M \times N$ all zero matrix is denoted by $0_{M\times N}$. For a vector u, we write $\|\mathrm{u}\|_2 = \sqrt{\mathrm{u}^{\mathsf{H}}\mathrm{u}}$ for its Euclidean norm and $\mathrm{supp}(\mathrm{u})$ for its support. For a set $\mathscr{S}$, $\mathrm{card}(\mathscr{S})$ denotes its cardinality. We use the notation $\mathrm{A}_{\mathscr{S}}$ to indicate the column submatrix of A consisting of the columns indexed by $\mathscr{S}$. Furthermore, $\|\mathrm{u}\|_0 \triangleq \mathrm{card}(\mathrm{supp}(\mathrm{u}))$ and $\|\mathrm{u}\|_1 = |u_0| + |u_1| + \ldots + |u_{N-1}|$ denote the ℓ_0-norm and the ℓ_1-norm of the vector $\mathrm{u} = [u_0, u_1, \cdots, u_{N-1}]^{\mathsf{T}}$, respectively. The value $\|\mathrm{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\mathrm{A}^{\mathsf{H}}\mathrm{A})}$ with $\mathrm{tr}(\cdot)$ being the trace of a matrix is the Froebenius norm of the matrix A, corresponding to the energy norm of its entries.

For the audio inpainting specific notation we adopt most of the conventions from [8, 9]: Let $\mathrm{x} \in \mathbb{R}^N$ be the time-domain signal and let the indices of its missing (or unreliable) samples be known. This will be denoted as the *gap*. The rest of the samples will be considered and called *reliable*. It is natural that the recovered signal should maintain consistency with the reliable part. To formally describe this, we introduce a (convex) set Γ_{x} as the set of all feasible signals

$$\Gamma_{\mathrm{x}} \triangleq \left\{ \mathrm{y} \in \mathbb{R}^N : \mathrm{M}_{\mathrm{R}}\mathrm{y} = \mathrm{M}_{\mathrm{R}}\mathrm{x} \right\}, \tag{1}$$

where $\mathrm{M}_{\mathrm{R}} : \mathbb{R}^N \to \mathbb{R}^N$ is the binary "reliable mask" projection operator keeping the signal samples corresponding to the reliable part, while setting the others to zero.

[1] Strictly speaking, these mathematical objects are not norms, but we adopt the common convention to refer to them as norms.

2 Gabor Systems and Frames

The audio inpainting approach presented in this contribution is based on the approximate time-frequency sparsity of audio signals. As the sparsifying transform, we apply the Gabor transform, also known as the Short-Time Fourier Transform (STFT), and – later on – a sparsity-optimized transform. The Gabor transform uses a single prototype window function g which is translated in time and modulated in frequency [11, 12]. The window serves to localize the signal in time. For the discrete Gabor transform (DGT), the translation of the window is characterized by the integer window shift (i.e., hop size) a. The number of modulations is denoted by M and will be referred to as the number of frequency channels. In the implementation this corresponds to the length of the fast Fourier transform (FFT). It is natural to require that the signal length N is divisible by a. Then, the system consists of $P = MN/a$ Gabor vectors $\mathrm{g}^{(p)} \in \mathbb{C}^N$, $p = 0,\ldots,P-1$. We will refer to as these vectors as the *Gabor atoms* and to the whole system $\{\mathrm{g}^{(p)} : p = 0,\ldots,P-1\} = \{\mathrm{g}^{(k,m)} : k = 0,\ldots,(N/a)-1, m = 0,\ldots,M-1\}$ with $\mathrm{g}^{(k,m)}_n = \mathrm{g}_{[n-ak]_N}\, e^{2\pi \mathrm{i}(n-ak)m/M}$ and $p = kM+m$ as the *Gabor dictionary*.

Note that the Gabor window g is usually identified with its shorter counterpart comprising only those elements of g that are within the smallest interval containing the support of g. The length of this interval is usually much smaller than N (and even smaller than M) and is denoted as *window length* w_{g}.

There exist suitable combinations of g and the parameters a and M such that the resulting Gabor system forms a frame for $\mathbb{C}^N$, i.e., any $\mathrm{x} \in \mathbb{C}^N$ can be represented in a stable way as a linear combination of the Gabor vectors allowing perfect reconstruction [11, 13]. Although Gabor bases can be constructed, they have undesired properties [12]. Therefore, overcomplete systems which allow non-unique signal representations

are usually preferred. Note the treatment of the signal as a complex vector, although we work only with real-valued audio signals.

In frame theory, the so-called *analysis operator* $\mathrm{A}:\mathbb{C}^N\to\mathbb{C}^P$ produces coefficients from the signal, whereas its adjoint $\mathrm{D}=\mathrm{A}^{\mathsf{H}}$, the *synthesis operator* $\mathrm{D}:\mathbb{C}^P\to\mathbb{C}^N$, generates a signal from the coefficients. Its composition $\mathrm{S}=\mathrm{DA}$ is denoted as *frame operator*. Whenever we deal with Gabor frames, we use the subscript notation $\mathrm{A_G}$ and $\mathrm{D_G}$ for analysis and synthesis operators, respectively, to emphasize the Gabor structure.

In this contribution, we restrict ourselves to frames, which correspond to the so-called *painless case* [14, 15]. This is the case nearly exclusively used in audio practice and is guaranteed if the length of the FFT is larger or equal to the window length, i.e., $M\geq w_{\mathrm{g}}$, see, e.g., [15]. Such frames have convenient properties from both theoretical and practical prospective and are typically the ones considered in signal processing. Their analysis, synthesis, and frame operators satisfy the "painless condition"

$$\mathrm{S}=\mathrm{A}^{\mathsf{H}}\mathrm{A}=\mathrm{DD}^{\mathsf{H}}=\begin{bmatrix}\mathrm{S}_{0,0} & & 0\\ & \ddots & \\ 0 & & \mathrm{S}_{N-1,N-1}\end{bmatrix} \quad (2)$$

with $\mathrm{S}_{n,n}>0$, $n=0,\ldots,N-1$, i.e., the frame operator matrix[2] is a diagonal matrix with diagonal elements strictly larger than zero.

3 SPAIN (Sparse Audio Inpainter)

In this section, we give a brief introduction to the SPAIN algorithm presented in [8]. As already mentioned, it is an adaptation of the SParse Audio DEclipper (SPADE) algorithm [10] to the inpainting problem. In fact, SPAIN and SPADE only differ in the definition of the set of feasible signals Γ_{x} (see (1) for its SPAIN definition). Moreover, SPAIN comes in two variants: the first one exploits analysis sparsity, the second one exploits synthesis sparsity (both in the time-frequency domain). Note that we restrict ourselves to the analysis variant in this contribution; for the synthesis variant we refer to the (full) journal version of this contribution [16].

[2]For notational simplicity, we do not distinguish between operators and their matrix representation with respect to the canonical basis throughout the paper.

The analysis variant of SPAIN aims at solving the following optimization task,

$$\min_{\mathrm{b,y}}\|\mathrm{b}\|_0 \quad \text{s.t.} \quad \mathrm{y}\in\Gamma_{\mathrm{x}} \quad \text{and} \quad \|\mathrm{Ay}-\mathrm{b}\|_2\leq\varepsilon, \quad (3)$$

where the minimizing y will be the reconstructed time-domain signal. It is obvious, that a brute-force solution of (3) will be infeasible due to its huge computational complexity. As an alternative, SPAIN applies the Alternating Direction Method of Multipliers (ADMM) [17] – carefully adapted – to the optimization of the above non-convex problem.

Of special importance in SPAIN is also the segment-wise application of the algorithms, where, first, the time-domain signal is segmented using overlapping windows, second, the algorithms are applied segment-wise using an overcomplete DFT frame, and finally, the restored blocks are combined via an overlap-add scheme.

If we would apply (3) to the signal consisting of the gap plus the adjacent parts of length w_{g} before and after the gap *as a whole* using a Gabor frame of appropriate dimension, the method would fail completely for medium length gaps. This is caused by the chosen regularizer $\|\cdot\|_0$, which penalizes the Gabor coefficients *globally* instead of *locally*. In particular, the reconstructed signal will be set to zero within the gap because its Gabor coefficients remain to be sparse globally. In order to cope with this behavior we propose a modification of the original SPAIN algorithm that works without segmentation.

4 The Modified SPAIN Algorithm

Let us recall, cf. Section 2, that Gabor systems impose a time-frequency structure. In case of the DGT, the $P=MN/a$ Gabor analysis coefficients can be rearranged[3] into an $M\times(N/a)$ matrix, whose column and row indices correspond to discrete time and discrete frequency, respectively. Mathematically, we express this *matrixification* via the (invertible) mapping $\tau:\mathbb{C}^P\to\mathbb{C}^{M\times(N/a)}$; its inverse mapping τ^{-1} represents a *vectorization*. Note further, that for any audio signal $\mathrm{x}\in\mathbb{R}^N$ the complex-valued matrix $\tau(\mathrm{A_G x})$ is conjugate-symmetric with respect to the frequency/row index. Hence, we do not loose any information if we only keep the rows corresponding

[3]Note that we have assumed that a divides N.

to the first $M' = \lfloor M/2 \rfloor + 1$ frequency indices; the remaining rows can be easily reobtained by exploiting conjugate-symmetry. We express this restriction operation via the mapping $\sigma : \mathbb{C}^{M\times(N/a)} \to \mathbb{C}^{M'\times(N/a)}$; its reversed operation, i.e., the conjugate-symmetric extension, is denoted by $\sigma^\dagger : \mathbb{C}^{M'\times(N/a)} \to \mathbb{C}^{M\times(N/a)}$. For notational simplicity, we also define the composite mappings $\eta(\cdot) \triangleq \sigma(\tau(\cdot))$ and $\eta^\dagger(\cdot) \triangleq \tau^{-1}(\sigma^\dagger(\cdot))$. The mapping $\mathbb{R}^N \to \mathbb{C}^{M'\times(N/a)}$, $\mathrm{x} \mapsto \eta\left(\mathrm{A_G x}\right)$ is often referred to as *real DGT* [18, 19].

For each $\mathrm{X} = [\mathrm{x}_0\, \mathrm{x}_1 \cdots \mathrm{x}_{(N/a)-1}] \in \mathbb{C}^{M'\times(N/a)}$ let us define the $\ell_{0,\infty}$-norm according to, cf. also [20],

$$\left\|\mathrm{X}\right\|_{0,\infty} \triangleq \max\left\{\left\|\mathrm{x}_0\right\|_0, \left\|\mathrm{x}_1\right\|_0, \ldots, \left\|\mathrm{x}_{(N/a)-1}\right\|_0\right\}. \quad (4)$$

Instead of (3), we propose to solve the following optimization task,

$$\min_{\mathrm{B,y}} \left\|\mathrm{B}\right\|_{0,\infty} \text{ s.t. } \mathrm{y} \in \Gamma_\mathrm{x} \text{ and } \left\|\eta\left(\mathrm{A_G y}\right) - \mathrm{B}\right\|_\mathrm{F} \leq \varepsilon. \quad (5)$$

Let us fix a sparsity parameter k. As shown in [16], the update rules of ADMM yield,

$$\mathrm{B}^{(i+1)} = \underset{\mathrm{B}:\|\mathrm{B}\|_{0,\infty}\leq k}{\arg\min} \left\|\mathrm{A_G y}^{(i)} - \eta^\dagger\left(\mathrm{B}\right) + \mathrm{r}^{(i)}\right\|_2 \quad (6a)$$

$$\mathrm{y}^{(i+1)} = \underset{\mathrm{y}\in\Gamma_\mathrm{x}}{\arg\min} \left\|\mathrm{A_G y} - \eta^\dagger\left(\mathrm{B}^{(i+1)}\right) + \mathrm{r}^{(i)}\right\|_2 \quad (6b)$$

$$\mathrm{r}^{(i+1)} = \mathrm{r}^{(i)} + \mathrm{A_G y}^{(i+1)} - \eta^\dagger\left(\mathrm{B}^{(i+1)}\right). \quad (6c)$$

Note that the conjugate-symmetry is preserved in each of these steps, provided that the initialization vectors $\mathrm{y}^{(0)}$ and $\mathrm{r}^{(0)}$ are real-valued and conjugate-symmetric, respectively.

The B-update (6a) is solved exactly [7, p. 42] by means of a time-frequency hard-thresholding operator $\mathscr{H}_k^{\mathrm{TF}} : \mathbb{C}^{M'\times(N/a)} \to \mathbb{C}^{M'\times(N/a)}$ defined according to

$$\mathscr{H}_k^{\mathrm{TF}}\left(\left[\mathrm{x}_0\, \mathrm{x}_1 \cdots \mathrm{x}_{(N/a)-1}\right]\right) = \left[\mathscr{H}_k(\mathrm{x}_0)\, \mathscr{H}_k(\mathrm{x}_1) \cdots \mathscr{H}_k(\mathrm{x}_{(N/a)-1})\right],$$

where $\mathscr{H}_k(\cdot)$ is essentially[4] the conventional hard-thresholding operator for vectors, which preserves the k elements with largest modulus and sets everything else to zero. More specifically, the B-update (6a) is given by

$$\mathrm{B}^{(i+1)} = \mathscr{H}_k^{\mathrm{TF}}\left(\eta\left(\mathrm{A_G y}^{(i)} + \mathrm{r}^{(i)}\right)\right).$$

As it is shown in [16], the y-update (6b) is equivalent to applying the inverse DGT using the canonical dual window to

$$\eta^\dagger\left(\mathrm{B}^{(i+1)}\right) - \mathrm{r}^{(i)}$$

and then projecting it onto the set of feasible solutions Γ_x.

The overall algorithm is initialized with a certain predefined sparsity parameter $k = s$, that will be increased in every t^{th} iteration by s, until the stopping criterion is met. Note that usually $s = t = 1$; however, as in the original SPAIN algorithm, we allow for more flexible settings in order to speed up the algorithm ($s > 1$) or supporting its convergence ($t > 1$).

Remark 1. The modified SPAIN algorithm aims at minimizing the $\ell_{0,\infty}$-norm as defined in (4). Roughly speaking, it searches for a signal in the feasible set Γ_x such that its real DGT is as sparse as possible in frequency direction *for all* time instants. In particular, for time instants close to the gap borders, a sparse representation with respect to frequency prohibits the occurrence of peaks, which are typically visible in the real DGT of the gapped signal, cf. Fig. 1c. This also illustrates the problem with "global" ℓ_0-minimization according to (3), which does not sufficiently penalize the signal in these specific "local" areas. Hence, by means of this approach, we avoid that the signal is set to zero within the filled gap without using the segmentation/overlap-add technique of the original SPAIN implementation. Note that we thereby only exploit sparsity in frequency direction.

5 Dictionary Learning

As pointed out, SPAIN relies on the observation that real-world audio signals have approximately sparse representations with respect to Gabor dictionaries. One might think, whether other dictionaries admit even more sparse signal representations and, in turn, improve the audio inpainting capabilities, if applied instead of a Gabor dictionary. Clearly, it is not trivial at all to come

[4] In order to cope with the conjugate-symmetry and still select the correct elements, the modulus has to be scaled for at most two elements of the vector (i.e., at frequency 0 and, for even M, also at frequency $M/2$).

up with a dictionary, which yields better results than the Gabor dictionary.

The idea, we intent to follow, is to *learn* an optimized dictionary from the reliable signal parts around the gap aiming at a more sparse representation within the gap as well so that the inpainting performance is enhanced. Of course, this is based on the assumption that the optimum sparsifying dictionary does not change too fast over time. Furthermore, we need some additional reliable signal parts in the neighborhood of the gap to learn the dictionary, which requires some distance between adjacent gaps, in case we deal with more that one gap. In order to keep the learning effort low, we will not learn it from scratch; instead we propose to "deform" a given Gabor dictionary.

5.1 Dictionary Learning Framework

Our starting point is the modified SPAIN algorithm as discussed in the previous section. Motivated by (5), we aim at finding a frame satisfying the painless condition (2) with analysis operator A such that for the degraded signal x, Ax is as sparse as possible with respect to the $\ell_{0,\infty}$-norm in a neighborhood of the gap. More formally, we intend to solve

$$\min_{\mathrm{A}} \left\| \big(\eta(\mathrm{Ax})\big)_{\mathscr{N}} \right\|_{0,\infty}, \quad \text{s.t.} \quad \mathrm{A}^{\mathsf{H}}\mathrm{A} \text{ is diagonal}, \quad (7)$$

where $\mathscr{N}$ represents the neighborhood. Let $\mathrm{A_G}$ denote the analysis operator of a given Gabor frame satisfying the painless condition (2). We next "deform" the Gabor frame using a unitary "deformation" operator $\mathrm{W}:\mathbb{C}^P \to \mathbb{C}^P$ by defining a modified analysis operator

$$\mathrm{A} = \mathrm{W}\mathrm{A_G}. \quad (8)$$

Note that we restrict to unitary deformations, since these preserve the painless condition (2) due to $\mathrm{A}^{\mathsf{H}}\mathrm{A} = \mathrm{A_G^{\mathsf{H}}}\mathrm{W}^{\mathsf{H}}\mathrm{W}\mathrm{A_G} = \mathrm{A_G^{\mathsf{H}}}\mathrm{A_G} = \mathrm{S_G}$. Furthermore, we are only interested in deformation operators, which increase sparsity in frequency direction and which preserve the conjugate-symmetry of the DGT. Therefore, we additionally impose the following structure on W,

$$\mathrm{W}:\mathbb{C}^P \to \mathbb{C}^P, \quad \mathrm{z} \mapsto \tau^{-1}\big(\mathrm{V}\,\tau(\mathrm{z})\big), \quad (9)$$

where[5] $\mathrm{V} \in \mathbb{C}^{M\times M}$ is a unitary matrix with the special form described below. Clearly, (9) represents a well-defined unitary operator. Regarding the structure of V, we have to distinguish between M even and M odd. However, for ease of exposition, we restrict ourselves to M even; for M odd, we refer to [16]. Let

$$\mathrm{V} = \mathrm{V_e} = \begin{bmatrix} 1 & & & 0 \\ & \mathrm{U} & & \\ & & 1 & \\ 0 & & & \mathrm{U}^{\mathsf{H}} \end{bmatrix} \quad (10)$$

with unitary $\mathrm{U} \in \mathbb{C}^{(M/2-1)\times(M/2-1)}$. It is not difficult to see that for a conjugate-symmetric z,

$$\mathrm{W}(\mathrm{z}) = \eta^{\dagger}\big(\mathrm{U_e}\,\eta(\mathrm{z})\big), \quad \mathrm{U_e} = \begin{bmatrix} 1 & & 0 \\ & \mathrm{U} & \\ 0 & & 1 \end{bmatrix}. \quad (11)$$

Combining (8), (11), and (7), we obtain

$$\min_{\mathrm{U_e}\in\mathscr{U}_{\mathrm{e}}} \left\| \mathrm{U_e}\big(\eta(\mathrm{A_G x})\big)_{\mathscr{N}} \right\|_{0,\infty},$$

where $\mathscr{U}_{\mathrm{e}}$ denotes the set of all unitary matrices $\mathrm{U_e}$ with the structure described in (11). Note, however, that this is a highly non-convex problem, so that we can only hope to find an approximate solution.

To this end, we follow the conventional approach to *relax* ℓ_0-norms to ℓ_1-norms [7, 21]. Furthermore, we replace the max-operation in the definition of the $\ell_{0,\infty}$-norm by a summation (i.e., in total we obtain an $\ell_{1,1}$-norm). This leads to the following optimization problem,

$$\hat{\mathrm{U}}_{\mathrm{e}} = \arg\min_{\mathrm{U_e}\in\mathscr{U}_{\mathrm{e}}} \sum_{q\in\mathscr{N}} \big\| \mathrm{U_e}\big(\eta(\mathrm{A_G x})\big)_q \big\|_1. \quad (12)$$

Remark 2. We note that the second replacement, i.e., max to summation, could be omitted. We are aware that retaining the max-operation would be desirable from a conceptual point of view, but in our implementations the summation variant was significantly more efficient in terms of computational complexity. In fact, the replacement even turned out to be mandatory because the computational load required by the max variant exceeded the capabilities of our simulation framework. Noteworthy, the summation variant searches for the optimum deformation matrix W such that the sparsity with respect to frequency is maximized[6] *on average* for all time instants in the considered neighborhood $\mathscr{N}$. Clearly, this approach does not guarantee that the sparsity is enhanced equally for all time instants in $\mathscr{N}$ but

[5] To avoid any confusion, $\mathrm{V}\,\tau(\mathrm{z})$ is a conventional matrix-matrix product.

[6] To avoid any confusion with this statement: we mean that the number of non-zeros is minimized, if we maximize sparsity.

for time instants within and around the gap (used for inpainting) we can still expect a considerably improved sparsity of the original signal. We also note that further restrictions in the basis optimization technique (see below and [16]) prohibit that the deformation matrix W differs too significantly from the identity matrix, so that stronger sparsity variations over time induced by W seem to be unlikely. This is also confirmed by our numerical experiments, see Section 6. Hence, for the learning phase, an average optimality criterion seems to be reasonable.

Finally, note that the minimization (12) is effectively carried out over the set $\mathscr{U}$ of unitary $(M/2-1)\times(M/2-1)$ matrices U according to (11), and that the optimum $\hat{\mathrm{U}}$ is given by

$$\hat{\mathrm{U}} = \arg\min_{\mathrm{U}\in\mathscr{U}} \sum_{q\in\mathscr{N}} \left\| \mathrm{U}\mathrm{E}_\mathrm{e}\big(\eta(\mathrm{A}_\mathrm{G}\mathrm{x})\big)_q \right\|_1, \qquad (13)$$

where

$$\mathrm{E}_\mathrm{e} \;=\; \left[0_{(M/2-1)\times 1}\;\; \mathrm{I}_{M/2-1}\;\; 0_{(M/2-1)\times 1}\right].$$

In order to solve (13), we resort to a basis optimization technique originally developed in the context of channel estimation [22, 23]. Note that standard convex optimization techniques cannot be used because the minimization problem (13) is non-convex (since $\mathscr{U}$ is not a convex set). For a detailed description of the used basis optimization technique we refer to [16].

Suppose we have found an optimized unitary matrix $\hat{\mathrm{U}}$. Inserting it into (11) and (10) yields $\hat{\mathrm{U}}_\mathrm{e}$ and $\hat{\mathrm{V}}_\mathrm{e}$, respectively, and furthermore, via (9) and (8), we obtain the analysis operator of the deformed frame as,

$$\hat{\mathrm{A}} : \mathbb{C}^N \to \mathbb{C}^P, \quad \mathrm{y} \mapsto \tau^{-1}\big(\hat{\mathrm{V}}_\mathrm{e}\,\tau(\mathrm{A}_\mathrm{G}\mathrm{y})\big).$$

With A_G replaced by $\hat{\mathrm{A}}$ we now apply the algorithm described in Sec. 4 in a way such that the learned sparsity-optimized frame is used instead of the Gabor frame.

5.2 Choosing the Learning Neighborhood $\mathscr{N}$

It remains to address the question, how to select the neighborhood $\mathscr{N}$ of the gap, which specifies the training matrix $\big(\eta(\mathrm{A}_\mathrm{G}\mathrm{x})\big)_\mathscr{N}$. It seems to be obvious that it should be as close as possible to the gap since this increases the chance that the learned sparse representation is also a sparse representation for the signal part within the gap. On the other hand, we also expect small "guard" intervals between the gap and training borders to be mandatory to add, in order to avoid that unreliable data from the gap influences our training data. According to the underlying Gabor structure, this guard interval should be at least of length w_g (length of Gabor window). Regarding the size of the neighborhood we have to somehow rely on intuition. A larger neighborhood yields a larger training matrix $\big(\eta(\mathrm{A}_\mathrm{G}\mathrm{x})\big)_\mathscr{N}$ and, probably, more accurate training results but it also increases the likelihood that signal variations occur within the neighborhood. Such signal variations could yield a dictionary adapted to signal parts, which are essentially independent of the signal within the gap. Suppose the signal $\mathrm{x}\in\mathbb{R}^N$ has a gap at the indices $n = n_\mathrm{B}, n_\mathrm{B}+1, \ldots, n_\mathrm{E}$. Taking into account the guard intervals, the neighborhood is given by $\mathscr{N} = \mathscr{N}_\mathrm{B}\cup\mathscr{N}_\mathrm{E}$ with[7]

$$\mathscr{N}_\mathrm{B} = \big\{k : \lfloor (n_\mathrm{B}-w_\mathrm{g})/a\rfloor - L_\mathscr{N} \le k < \lfloor (n_\mathrm{B}-w_\mathrm{g})/a\rfloor\big\},$$
$$\mathscr{N}_\mathrm{E} = \big\{k : \lceil (n_\mathrm{E}+w_\mathrm{g})/a\rceil < k \le \lceil (n_\mathrm{E}+w_\mathrm{g})/a\rceil + L_\mathscr{N}\big\}$$

i.e., a part before and a part after the gap. Its overall length is $2L_\mathscr{N}$. We will address the influence of the length parameter $L_\mathscr{N}$ by means of numerical experiments, cf. Subsec. 6.1.

6 Simulation Results

This section presents a numerical evaluation of our dictionary learning technique for sparse audio inpainting. As main performance measure, we use the signal-to-noise ratio (SNR), defined as [1]

$$\mathrm{SNR}(\mathrm{x}_\mathrm{orig}, \mathrm{x}_\mathrm{inp}) = 10\log_{10}\frac{\|\mathrm{x}_\mathrm{orig}\|_2^2}{\|\mathrm{x}_\mathrm{orig}-\mathrm{x}_\mathrm{inp}\|_2^2} \quad [\mathrm{dB}],$$

where x_orig and x_inp denote original and inpainted signal within the gaps, respectively. Clearly, higher SNR values reflect better reconstruction. We compute the average SNR by first computing the particular values of SNR in dB, and then taking the average. Furthermore, we also include the PEMO-Q criterion [24], which takes a model of the human auditory system into account. Therefore, it is closer to the subjective evaluation than the SNR. The measured quantity called *objective difference grade* (ODG) can be interpreted as the degree of perceptual similarity of x_orig and x_inp. The ODG attains values from -4 (very annoying) up to 0 (imperceptible), expressing the effect of audio artifacts in the restored signal.

We use a collection of ten music recordings sampled at 44.1 kHz, with different levels of sparsity with respect

[7] We assume that the gap is sufficiently centered within {0,...,(N/a)-1}, so that $0 \le \lfloor (n_\mathrm{B}-w_\mathrm{g})/a\rfloor - L_\mathscr{N}$ and $\lceil (n_\mathrm{E}+w_\mathrm{g})/a\rceil + L_\mathscr{N} < N/a$.

to the original Gabor representation. Our signals were chosen from the EBU SQAM dataset [25]. In each test instance, the input was a signal containing 5 gaps at random positions. The lengths of the gaps ranged from 5ms up to 50ms. For fixed lengths, the results over all ten signals consisting of the 5 gaps were averaged.

Throughout our experiments we used a tight Gabor frame with the Hann window of length $w_g = 2800$ samples (approximately 64ms), window shift $a = 700$ samples and with M = 2800 frequency channels. We used the fast implementation of Gabor transforms offered by the LTFAT toolbox [18, 19] in our computations, and we adopted its time-frequency conventions. For the basis optimization algorithm, cf. [16], we set the maximum number of iterations to $r_{\max} = 20$, the off-diagonal parameter to $d = 3$, and the remaining parameters to $\rho_{\text{start}} = 1$ and $\varepsilon = 2^{-20}$. Finally, all SPAIN variants used the input parameters $s = t = 1$.

6.1 Sparsity with Respect to Learned Dictionary

The purpose of this subsection is to analyze the sparsity of the analysis coefficients of real-world audio signals with respect to the optimized frame and to compare it with the analysis coefficients using the original Gabor frame.

Fig. 1 illustrates the analysis coefficients of the signal "a35_glockenspiel" from the EBU SQAM dataset [25]. Subfigures (a) and (b) depict the signal without gap, (c) and (d) with gap of length 40ms. Subfigures (a) and (c) use the Gabor dictionary, (b) and (d) the learned dictionary. The coefficients within the green rectangles are used for training, corresponding to a neighborhood length parameter of $L_{\mathscr{N}} = 4w_g/a = 16$. As expected, the representations with respect to the learned dictionary are more sparse than the ones with respect to the Gabor dictionary. Although the training coefficients have a certain distance from the gap area, the learned dictionary also sparsifies the original signal within the gap area. This confirms our assumption that the sparsifying dictionary does not change too fast over time.

Moreover, we are interested how the size of the learning neighborhood $\mathscr{N}$ around the gap impacts the sparsity of the learned representation. To this end, Fig. 2 depicts the analysis coefficients of the signal "a25_harp" from the EBU SQAM dataset [25]. Subfigures (a) and (c) use Gabor dictionary, (b) and (d) use learned dictionaries. The coefficients within the green rectangles are used for training: subfigure (b) is obtained from training according to (a) and subfigure (d) is obtained from training according to (c). Note that the neighborhood length parameter used to obtain (b) was $L_{\mathscr{N}} = 4w_g/a = 16$, whereas the neighborhood length parameter used to obtain (d) was $L_{\mathscr{N}} = 2w_g/a = 8$. It is seen that the analysis coefficients with respect to the dictionary obtained by a larger training set are more sparse, especially for the coefficients in the training areas. However, within the relevant gap area, it is very difficult to notice any differences, so that a more informative comparison has to be done in terms of inpainting performance. Indeed, according to our systematic study in [16], the version with smaller learning neighborhood turned out to be superior to the one with larger neighborhood in the vast majority of instances, so that we restricted to the training neighborhood with length parameter $L_{\mathscr{N}} = 2w_g/a = 8$ for all further comparisons.

6.2 Performance Comparison

Here, we compare our proposed methods with other existing audio inpainting methods. In particular, we simulated the following audio inpainting techniques (the abbreviations in the rectangular brackets are used in Fig. 3: 'A-' stands for analysis variant):

- The modified SPAIN algorithm introduced in Sec. 4 using a Gabor dictionary [A-SPAIN-MOD].
- The modified SPAIN algorithm using a *learned* dictionary [A-SPAIN-LEARN].
- The original SPAIN algorithm presented in [8] using a frame-wise DFT dictionary with redundancy 4 [A-SPAIN].
- The frame-wise Janssen algorithm [4] with autoregressive model order $p = \min(3H+2, w_g/3)$, where H denotes the number of missing samples within the current frame (window), and the number of iterations was set to 50 [JANSSEN].

Fig. 3 shows the inpainting performance of the aforementioned algorithms after averaging over all ten signals with five gaps. It is seen that in terms of ODG values the proposed methods A-SPAIN-MOD and A-SPAIN-LEARNED outperform the other inpainting methods over the whole gap length range (5ms – 50ms). Among those two, A-SPAIN-LEARNED, which uses the learned dictionary, achieves best performance. We also observe that A-SPAIN-LEARNED is superior to

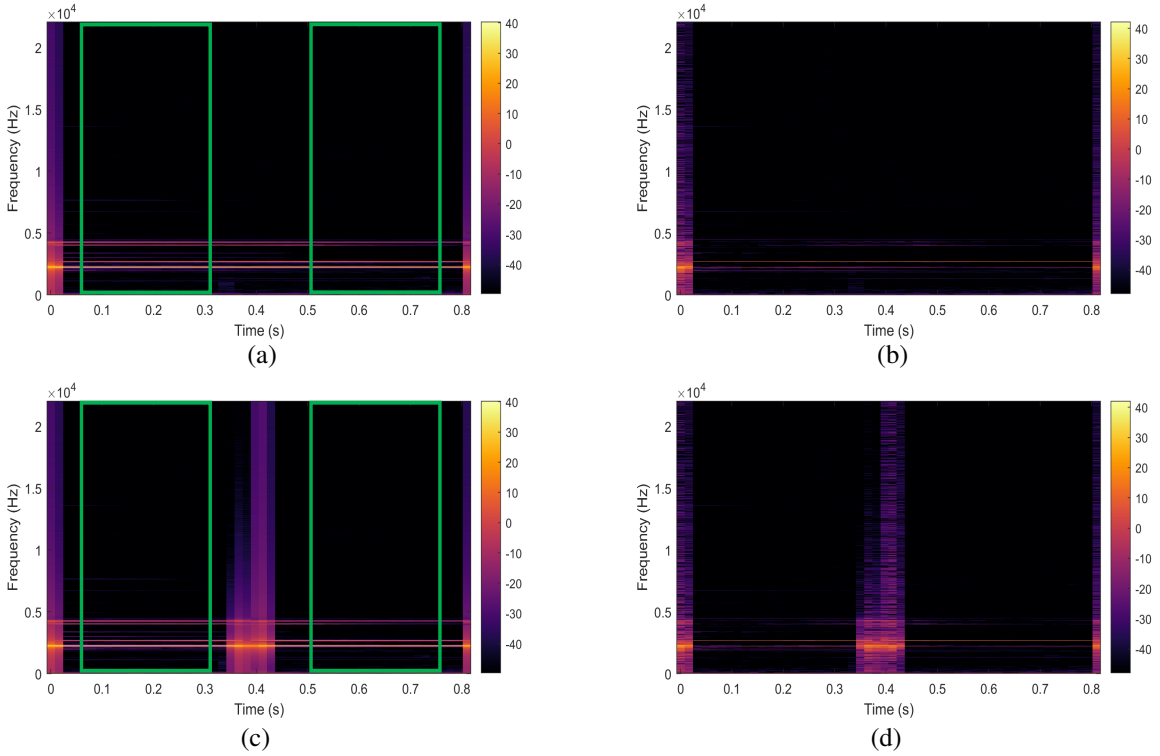


Fig. 1: Analysis coefficients of signal "a35_glockenspiel". (a) and (b) depict the signal without gap, (c) and (d) with gap of length 40ms. (a) and (c) use Gabor dictionary, (b) and (d) use learned dictionary. The coefficients within the green rectangles are used for training.

any other algorithm in terms of SNR, thus, illustrating the large benefit of a sparsity-optimized dictionary. Note, however, that this comes at cost of increased computational complexity due to the additional learning step.

7 Conclusions

We presented a dictionary learning framework for audio inpainting. Our proposed method learns the dictionary from reliable parts around the gap with the overall goal to obtain a signal representation with increased sparsity. Moreover, we presented a modified SPAIN algorithm which replaces the conventional hard thresholding operator by a specific time-frequency hard thresholding operator, so that the segment-wise processing of the original SPAIN algorithm can be avoided. Our experimental results demonstrated that the proposed dictionary learning approach yields large performance gains in combination with the modified SPAIN algorithm and significantly outperforms any other inpainting approach compared with.

References

[1] Adler, A., Emiya, V., Jafari, M., Elad, M., Gribonval, R., and Plumbley, M., "Audio Inpainting," *IEEE Trans. Audio, Speech, and Language Processing*, 20(3), pp. 922–932, 2012.

[2] Perkins, C., Hodson, O., and Hardman, V., "A survey of packet loss recovery techniques for streaming audio," *IEEE network*, 12(5), pp. 40–48, 1998.

[3] Godsill, S., Rayner, P., and Cappé, O., "Digital audio restoration," in *Applications of digital signal processing to audio and acoustics*, pp. 133–194, Springer, 2002.

[4] Janssen, A., Veldhuis, R., and Vries, L., "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Trans. Ac., Sp., Sign. Proc.*, 34(2), pp. 317–330, 1986.

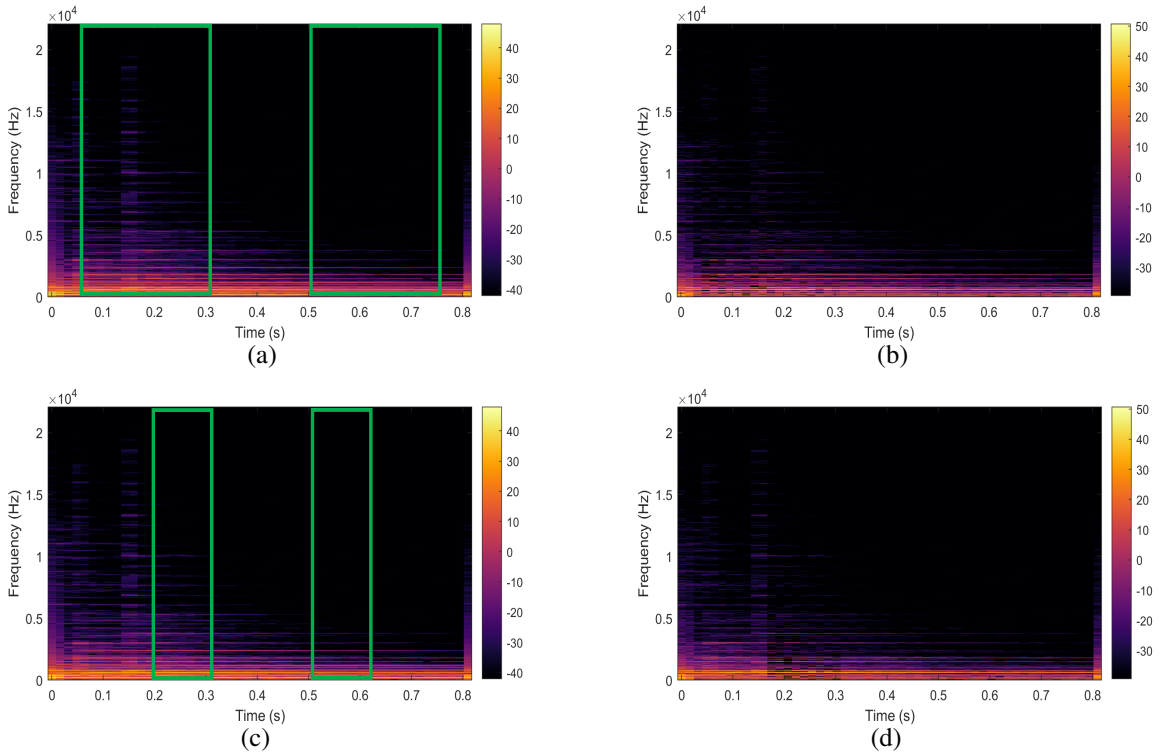


Fig. 2: Analysis coefficients of signal "a25_harp" without gap. (a) and (c) use Gabor dictionary, (b) and (d) use learned dictionary based on different training neighborhoods. The coefficients within the green rectangles are used for training: (b) is obtained from training according to (a) and (d) is obtained from training according to (c).

[5] Oudre, L., "Interpolation of missing samples in sound signals based on autoregressive modeling," *Image Processing On Line*, 8, pp. 329–344, 2018.

[6] Etter, W., "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Trans. Signal Process.*, 44(5), pp. 1124–1135, 1996.

[7] Foucart, S. and Rauhut, H., *A Mathematical Introduction to Compressive Sensing*, Applied and Numerical Harmonic Analysis, Birkhäuser, Basel, 2013.

[8] Mokrỳ, O., Záviška, P., Rajmic, P., and Veselỳ, V., "Introducing SPAIN (SParse Audio INpainter)," in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, IEEE, 2019.

[9] Mokrỳ, O. and Rajmic, P., "Audio Inpainting: Revisited and Reweighted," *arXiv preprint arXiv:2001.02480*, 2020.

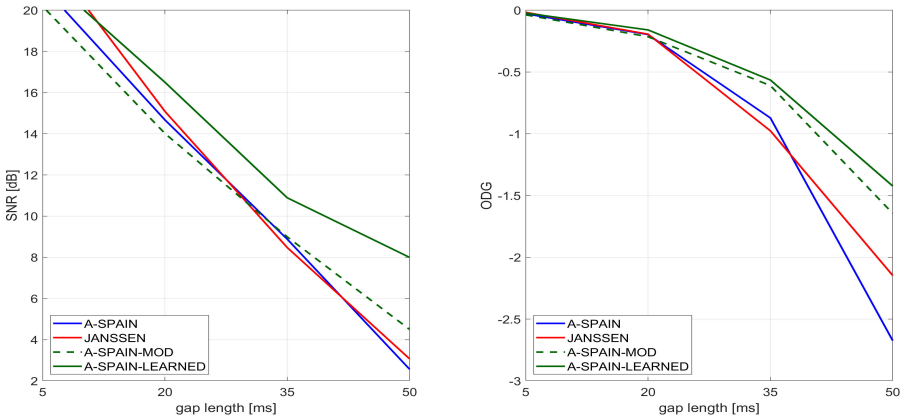
[10] Kitić, S., Bertin, N., and Gribonval, R., "Sparsity and cosparsity for audio declipping: a flexible non-convex approach," in *Int. Conf. o. Lat. Var. Anal. a. Sign. Sep.*, pp. 243–250, Springer, 2015.

[11] Christensen, O. et al., *An Introduction to Frames and Riesz Bases*, Springer, 2016.

[12] Gröchenig, K., *Foundations of Time-Frequency Analysis*, Birkhäuser, Boston, MA, 2001.

[13] Feichtinger, H. G. and Strohmer, T., *Advances in Gabor Analysis*, Springer Science & Business Media, 2012.

[14] Daubechies, I., Grossmann, A., and Meyer, Y., "Painless nonorthogonal expansions," *Journal of Mathematical Physics*, 27(5), pp. 1271–1283, 1986.



(a) (b)

Fig. 3: Overall performance comparison of various audio inpainting algorithms. (a) depicts results in terms of SNR, (b) in terms of ODG values.

[15] Balazs, P., Dörfler, M., Jaillet, F., Holighaus, N., and Velasco, G., "Theory, implementation and applications of nonstationary Gabor frames," *Journal of computational and applied mathematics*, 236(6), pp. 1481–1496, 2011.

[16] Tauböck, G., Rajbamshi, S., and Balazs, P., "Dictionary Learning for Sparse Audio Inpainting," 2020, submitted.

[17] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, 3(1), pp. 1–122, 2011.

[18] Søndergaard, P. L., Torrésani, B., and Balazs, P., "The linear time frequency analysis toolbox," *Int. J. Wavel., Multires. a. Inf. Process.*, 10(04), p. 1250032, 2012.

[19] Průša, Z., Søndergaard, P. L., Holighaus, N., Wiesmeyr, C., and Balazs, P., "The Large Time-Frequency Analysis Toolbox 2.0," in *Sound, Music, and Motion*, LNCS, pp. 419–442, Springer International Publishing, 2014.

[20] Balazs, P., Doerfler, M., Kowalski, M., and Torresani, B., "Adapted and adaptive linear time-frequency representations: a synthesis point of view," *IEEE Sig. Proc. Mag.*, 30(6), pp. 20–31, 2013.

[21] Donoho, D. L. and Elad, M., "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proceedings of the National Academy of Sciences*, 100(5), pp. 2197–2202, 2003.

[22] Tauböck, G. and Hlawatsch, F., "Compressed sensing based estimation of doubly selective channels using a sparsity-optimized basis expansion," in *2008 16th European Signal Processing Conference*, Lausanne, Switzerland, 2008.

[23] Tauböck, G., Hlawatsch, F., Eiwen, D., and Rauhut, H., "Compressive Estimation of Doubly Selective Channels in Multicarrier Systems: Leakage Effects and Sparsity-Enhancing Processing," *IEEE J. Sel. Topics Signal Process.*, 4(2), pp. 255–271, 2010.

[24] Huber, R. and Kollmeier, B., "PEMO-Q–A new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6), pp. 1902–1911, 2006.

[25] "EBU SQAM CD: Sound quality assessment material recordings for subjective tests," 2008, https://tech.ebu.ch/publications/sqamcd.