Dictionary Learning for Sparse Audio Inpainting

Georg Tauböck, Shristi Rajbamshi, and Peter Balazs

***Abstract*—The objective of audio inpainting is to fill a gap in an audio signal. This is ideally done by reconstructing the original signal or, at least, by inferring a meaningful surrogate signal. We propose a novel approach applying sparse modeling in the time-frequency (TF) domain. In particular, we devise a dictionary learning technique which learns the dictionary from reliable parts around the gap with the goal to obtain a signal representation with increased TF sparsity. This is based on a basis optimization technique to deform a given Gabor frame such that the sparsity of the analysis coefficients of the resulting frame is maximized. Furthermore, we modify the SParse Audio INpainter (SPAIN) for both the analysis and the synthesis model such that it is able to exploit the increased TF sparsity and—in turn—benefits from dictionary learning. Our experiments demonstrate that the developed methods achieve significant gains in terms of signal-to-distortion ratio (SDR) and objective difference grade (ODG) compared with several state-of-the-art audio inpainting techniques.**

***Index Terms*—Audio inpainting, convex, dictionary, frame, Gabor, learning, optimization, sparsity, time-frequency.**

I. Introduction

AUDIO signals are often prone to localized distortions resulting in modification or even loss of certain sections. A signal processing technique that aims at restoring such gaps, i.e., missing consecutive samples, while still keeping perceptible audio artifacts as small as possible is usually referred to as *audio inpainting* [1]. It has attracted a lot of attention as it has various important applications. Examples include reconstruction of audio samples caused by scratches in CDs or old recordings [2], compensation of audio packet losses in communication networks [3], [4], and others.

A typical approach to deal with the audio inpainting task is to utilize reliable signal parts jointly with some prior information about the signal. Some of the first methods were proposed by Janssen *et al.* [5], [6] and Etter [7]. These techniques exploit available signal information in the original (time) domain and fill the missing samples by linear prediction using (learned) autoregressive coefficients. Due to their excellent performance they are still considered state-of-the-art. For a more comprehensive study on autoregressive-based audio inpainting we refer to, e.g., [8]–[10].

Since the arrival of sparse signal representations and compressive sensing [11]–[15], several other audio inpainting techniques have been introduced [1], [16]–[18], most notably [1], which coined the term "audio inpainting" motivated by analogous image processing tasks. These methods tackle the inpainting problem by leveraging the (approximate) sparsity of real-world audio signals with respect to suitable (redundant) dictionaries.

It is noteworthy that for *long* gaps (≥ 100 ms) all the aforementioned methods start to fail due to the non-stationarity of audio signals over longer periods of time. Therefore, other techniques have been introduced for long gaps, e.g., sinusoidal modeling [19], [20], similarity graph approaches [4], [21], or methods based on deep neural networks [22], [23]. Their focus is no longer to reconstruct the true original signal but rather to fill the gaps in a perceptually pleasant and meaningful way.

In this contribution, we concentrate on *medium* gap length settings (10 ms – 100 ms). Note that the design of audio inpainting methods for such scenarios is quite challenging since it is still intended to recover the original signal but the gap duration is close to the limit for which stationarity can be justified. This is probably the reason, why research contributions dealing with medium gaps are essentially only the few ones mentioned above. This is in stark contrast to the related problem of audio declipping with a substantial amount of available literature, e.g., [24]–[33]. Clipping is a non-linear distortion mechanism that degrades the signal whenever the modulus of its amplitude exceeds a certain threshold. The underlying clipping model allows to exploit signal information also within the degraded signal parts. Therefore, it allows for more reliable signal reconstructions. Note that inpainting with randomly selected missing samples as studied in, e.g., [16] is closer to the clipping problem. As long as the underlying probability model avoids occurrences of long sequences of missing samples with high probability, reconstruction error probabilities can be expected to be small since the *short* gap setting (≤ 10 ms) is present with high probability.

A. Motivation and Contributions

Our audio inpainting approach is inspired by the recently introduced algorithm *SParse Audio INpainter (SPAIN)* [17] and extends our work in [34]. SPAIN is an adaptation of the so-called SParse Audio DEclipper (SPADE) algorithm [28] to the inpainting problem and exploits sparsity with respect to tight (Gabor) frames [35], [36]. In [17], both synthesis and analysis models are discussed and an efficient implementation with segment-wise application of the algorithm is presented. More specifically, the time-domain signal is segmented using overlapping Hann windows, sparsity with respect to an overcomplete

Manuscript received May 8, 2020; revised September 4, 2020, October 28, 2020, and December 12, 2020; accepted December 14, 2020. Date of publication December 21, 2020; date of current version January 29, 2021. This work was supported in part under WWTF Grant INSIGHT (MA16-053), and in part under FWF Grant FLAME (Y 551-N13). Parts of this work have been previously published at the Audio Engineering Society Convention 149, Online, Oct. 2020. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Pavel Rajmic. *(Corresponding author: Georg Tauböck.)*

The authors are with the Acoustics Research Institute, Austrian Academy of Sciences, A-1040 Vienna, Austria (e-mail: georg.tauboeck@oeaw.ac.at; shristi.rajbamshi@oeaw.ac.at; peter.balazs@oeaw.ac.at).

Digital Object Identifier 10.1109/JSTSP.2020.3046422

Discrete Fourier Transform (DFT) dictionary is exploited, and the restored blocks are combined using an overlap-add scheme. However, like many other sparsity-based inpainting techniques, SPAIN suffers from a drop of the signal's energy withing the filled gap, which is the more pronounced the larger the gap is. As explained in [18], this is caused by the chosen regularizer that penalizes the deviation of transformed signal's coefficients from the model under consideration. While [18] investigates weighted ℓ_1-norms to cope with these systematic biases, we propose a modification of the original SPAIN algorithm (for both analysis and synthesis variants): instead of the mentioned segment-wise processing scheme, we apply the algorithm to the signal parts in the neighborhood of the gap as a whole. Furthermore, we replace the involved ℓ_0-norm[1] by an $\ell_{0,\infty}$-norm[1], where the supremum is taken over time.

Finally, and this is certainly our key contribution, we present a basis optimization technique which modifies the underlying dictionary in order to obtain a representation with increased sparsity. Here, the main idea is to learn the dictionary from reliable signal parts around the gap. By means of the resulting sparsity enhanced dictionary our modified SPAIN algorithm exhibits a significantly improved reconstruction performance. Note that we are not aware of any other dictionary learning techniques for audio inpainting for the medium gap setting. There are several related dictionary learning contributions for audio inpainting but their focus is on short gaps or on audio declipping [30], [37], [38]. We also want to emphasize that we cannot simply use out-of-the-shelf dictionary learning methods like K-SVD [39], MOD [40], or others [41], because we require our learned dictionary to satisfy specific structural properties, see Subsection V-A.

B. Notation

Scalars, vectors, and matrices are designated by Roman letters $a, b, \ldots$, $\mathrm{a}, \mathrm{b}, \ldots$, and $\mathrm{A}, \mathrm{B}, \ldots$, respectively. 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 $^\mathsf{T}$, $^\mathsf{H}$, and * denote transposition, Hermitian transposition, and (element-wise) complex conjugation, respectively. I_N stands for the $N \times N$ identity matrix; $0_{M\times N}$ stands for the $M \times N$ all zero matrix. The floor function $\lfloor a \rfloor$ is defined as the largest integer $\leq a$, whereas $[\,\cdot\,]_N = [\,\cdot \mod N]$ abbreviates the modulo-N operation due to circular indexing. For a set $\mathcal{S}$, we write $\mathrm{card}(\mathcal{S})$ for its cardinality and $\chi_{\mathcal{S}}(\cdot)$ is the indicator function on $\mathcal{S}$, which is 0 if its argument is in $\mathcal{S}$ and ∞ otherwise. The notation $\mathrm{A}_{\mathcal{S}}$ is used to indicate the column submatrix of A consisting of the columns indexed by $\mathcal{S}$. Similarly, for $\mathrm{x} \in \mathbb{C}^N$, $\mathrm{x}_{\mathcal{S}}$ denotes the subvector in $\mathbb{C}^{\mathrm{card}(\mathcal{S})}$ consisting of the entries of x indexed by $\mathcal{S}$. For a vector $\mathrm{u} = [u_0, u_1, \ldots, u_{N-1}]^\mathsf{T}$, supp(u) denotes its support, i.e., the set where the coefficients are non-zero, and $\|\mathrm{u}\|_0 = \mathrm{card}(\mathrm{supp}(\mathrm{u}))$, $\|\mathrm{u}\|_1 = |u_0| + |u_1| + \ldots + |u_{N-1}|$, and $\|\mathrm{u}\|_2 = \sqrt{\mathrm{u}^\mathsf{H}\mathrm{u}}$, are its ℓ_0-norm, ℓ_1-norm, and ℓ_2-norm, respectively. For a matrix A, $\mathrm{tr}(\mathrm{A})$ is its trace, $\|\mathrm{A}\|_\mathrm{F} = \sqrt{\mathrm{tr}(\mathrm{A}^\mathsf{H}\mathrm{A})}$ is its Froebenius norm, and $\|\mathrm{A}\|_{\infty,\infty}$ is the largest modulus of all entries of A. $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote real and imaginary part of its argument, respectively.

For the audio inpainting specific notation we adopt most of the conventions from [17], [18]: Let $\mathrm{x} \in \mathbb{R}^N$ be the time-domain signal and assume that the indices of its missing (or unreliable) samples are known. This will be referred to as the *gap*. The samples outside the gap will be considered and called *reliable*. Clearly, the recovered signal should maintain consistency with the reliable part. In order to mathematically describe this, we introduce a (convex) set Γ_x as the set of all feasible signals

$$\Gamma_\mathrm{x} \triangleq \left\{\mathrm{y} \in \mathbb{R}^N : \mathrm{M_R}\mathrm{y} = \mathrm{M_R}\mathrm{x}\right\}, \qquad (1)$$

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

II. Gabor Systems and Frames

The audio inpainting approach presented in this contribution relies on the observation that audio signals are (approximately) sparse in the time-frequency domain. More specifically, the Gabor transform—also denoted as Short-Time Fourier Transform (STFT)—of a real-world audio signal typically distributes main portions of the signal's energy only within some subareas of the time-frequency plane; the remaining areas contain merely small fractions of the signal energy. Note that the Gabor transform computes inner products of the input signal with time-shifted and frequency-modulated window functions [35], [36]. For the discrete Gabor transform (DGT), the integer-valued hop size a specifies the time-translations of the window g. The number[2] of modulations (frequency shifts) is denoted by M, so that there are in total M frequency channels. It is natural to require that a divides the signal length N. Then, the system consists of $P = MN/a$ *Gabor atoms* $\mathrm{g}^{(p)} \in \mathbb{C}^N$, $p = 0, \ldots, P-1$. The whole system

$$\begin{aligned}&\{\mathrm{g}^{(p)} : p = 0, \ldots, P-1\}\\ &\quad = \{\mathrm{g}^{(k,m)} : k = 0, \ldots, (N/a)-1, m = 0, \ldots, M-1\}\end{aligned}$$

with

$$\mathrm{g}_n^{(k,m)} = \mathrm{g}_{[n-ak]_N}\, \mathrm{e}^{2\pi\mathrm{i}(n-ak)m/M} \quad \text{and} \quad p = kM + m$$

is referred to as the *Gabor dictionary*.

Note that the Gabor window g is usually identified with its shorter counterpart keeping only those elements of g that are within the smallest interval containing the support of g. The length of this interval is denoted as *window length* w_g and is usually much smaller than N.

For suitable combinations of g and the parameters a and M, the resulting Gabor system forms a frame for $\mathbb{C}^N$ and, hence, allows perfect reconstruction [35], [42]. That is, any $\mathrm{x} \in \mathbb{C}^N$ can be represented in a stable way as a linear combination of the Gabor vectors. Although Gabor bases can be constructed, they have undesired properties [36]. Therefore, overcomplete

[1] We adopt the common convention to refer to these mathematical objects as norms, although they are not norms in the strict sense.

[2] In implementations, this value corresponds to the length of the fast Fourier transform (FFT).

systems which allow non-unique signal representations are usually preferred. Note that the Gabor atoms are complex vectors, although we work only with real-valued audio signals.

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

In this paper, we focus on frames which correspond to the so-called *painless case* [36], [43], [44]. Such frames are convenient from both theoretical and practical perspectives, since they allow for a simple and computationally efficient computation of their canonical dual frames [35], and are typically the ones considered in signal processing. Their analysis, synthesis, and frame operators satisfy the "painless condition"

$$\mathrm{S} = \mathrm{A}^{\mathsf{H}}\mathrm{A} = \mathrm{D}\mathrm{D}^{\mathsf{H}} = \begin{bmatrix} \mathrm{S}_{0,0} & & 0 \\ & \ddots & \\ 0 & & \mathrm{S}_{N-1,N-1} \end{bmatrix} \tag{2}$$

with $\mathrm{S}_{n,n} > 0$, $n = 0, \ldots, N-1$, i.e., the frame operator matrix[3] is a diagonal positive-definite matrix. Clearly, tight frames, i.e., frames where $\mathrm{S} = \lambda \mathrm{I}_N$, fall within the class of painless frames. For Gabor frames, $w_{\mathrm{g}} \leq M$ implies (2), see, e.g., [44, Corollary 3]. We note that [33] considers the painless case in the context of audio declipping (but finally only uses the tight setting).

III. SPAIN (SPARSE AUDIO INPAINTER)

In this section, we will briefly introduce the SPAIN algorithm presented in [17]. Being an adaptation of the SParse Audio DEclipper (SPADE) algorithm [28] to the inpainting problem, SPAIN differs from SPADE only in the definition of the set of feasible signals Γ_{x} (see (1) for its SPAIN definition). Accordingly, 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 the sparse signal processing literature relied for a long time on the so-called synthesis model, where one seeks for a small number of coefficients synthesizing the desired signal [39], [45]–[48]. More recently, the analysis model appeared, where one looks directly for the signal, with the constraint that its coefficients after analysis are sparse [28], [49]. Note, however, that we can only expect these coefficients to be approximately sparse, since underlying uncertainty principles restrict the maximum degree of sparsity in this domain [36], [50], [51]. If the synthesis/analysis operators are invertible, both approaches are equivalent. This corresponds to the basis case, and as said above, usually this is avoided.

The two variants of SPAIN aim at solving the following optimization tasks,

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

$$\min_{\mathrm{b},\mathrm{y}} \|\mathrm{b}\|_0 \quad \text{s.t.} \quad \mathrm{y} \in \Gamma_{\mathrm{x}} \quad \text{and} \quad \|\mathrm{y} - \mathrm{Db}\|_2 \leq \epsilon, \tag{3b}$$

where (3a) and (3b) present the formulation referred to as the analysis and the synthesis variant, respectively. In both cases the y that (jointly with b) achieves the minimum will be the reconstructed time-domain signal. However, due to its huge computational complexity, a brute-force solution of (3) is infeasible. As an alternative, SPAIN applies the Alternating Direction Method of Multipliers (ADMM) [52]—carefully modified—to minimize the above non-convex problems.

The ADMM is able to solve problems of the form

$$\min_{\mathrm{y}} f(\mathrm{y}) + g(\mathrm{Ay}), \tag{4}$$

where $\mathrm{y} \in \mathbb{C}^N$, $\mathrm{A} : \mathbb{C}^N \to \mathbb{C}^P$ is a linear operator, and f and g are real convex functions. A reformulation of (4) yields

$$\min_{\mathrm{y},\mathrm{b}} f(\mathrm{y}) + g(\mathrm{b}) \quad \text{s.t.} \quad \mathrm{Ay} - \mathrm{b} = 0,$$

from which the Augmented Lagrangian (in scaled form) is computed. Finally, this Lagrangian is minimized individually with respect to each involved variable in an iterative fashion (also incorporating a combination update step) until a sufficiently accurate solution is obtained. Note that in case of SPAIN only an approximate solution can be expected, since (3) violates the convexity assumption.

Of crucial importance for SPAIN is also its segment-wise application, where, first, the time-domain signal is segmented using overlapping windows; second, the algorithm is applied to each segment individually using an overcomplete DFT frame; and, finally, the restored blocks are combined via an overlap-add scheme.

If we would refrain from the segment-wise implementation and would aim at solving (3) for 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, this approach would fail completely for medium length gaps. This behavior is caused by the chosen regularizer $\|\cdot\|_0$, which penalizes the Gabor coefficients *globally* instead of *locally*. More specifically, the reconstructed signal will vanish within the gap because the Gabor coefficients remain to be sparse globally.[4] As an alternative strategy that avoids the need for segmentation we propose a modification of the original SPAIN algorithm in the next section. Note that the segmentation step in the original SPAIN algorithm implicitly introduces *local* processing.

IV. MODIFIED SPAIN ALGORITHM

Let us recall, see Section II, that Gabor systems impose a time-frequency structure. In the discrete setting, the $P = MN/a$ analysis coefficients of the DGT can be rearranged[5] into an $M \times (N/a)$ matrix, whose column and row indices correspond to discrete time and discrete frequency, respectively. Mathematically, this *matrixification* can be expressed

[3] For ease of notation, we do not distinguish between operators and their matrix representation with respect to the canonical basis throughout the paper.

[4] Note that the reduced sparsity in frequency direction around the gap edges due to the introduced discontinuities will be compensated for by the increased sparsity in the middle of the gap.

[5] Note that we have assumed that N is a multiple of a.

via the (invertible) mapping $\tau : \mathbb{C}^{P} \to \mathbb{C}^{M\times(N/a)}$; its inverse mapping τ^{-1} represents a *vectorization*. Furthermore, for any real-valued audio signal x, the complex-valued matrix $\tau(\mathrm{A_G x})$ is conjugate-symmetric with respect to the frequency/row index. Thus, the available information is preserved if only the rows corresponding to the first $M' = \lfloor M/2 \rfloor + 1$ frequency indices are kept; the remaining rows are easily re-obtained by invoking conjugate-symmetry. We represent this restriction operation with 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 introduce the composite mappings $\eta(\cdot) \triangleq \sigma(\tau(\cdot))$ and $\eta^{\dagger}(\cdot) \triangleq \tau^{-1}(\sigma^{\dagger}(\cdot))$. The mapping $\mathbb{R}^{N} \to \mathbb{C}^{M'\times(N/a)}$, $\mathrm{x} \mapsto \eta(\mathrm{A_G x})$ is well known as *real DGT* [53], [54]. For each $\mathrm{X} = [\mathrm{x}_0\, \mathrm{x}_1 \cdots \mathrm{x}_{(N/a)-1}] \in \mathbb{C}^{M'\times(N/a)}$ let

$$\left\|\mathrm{X}\right\|_{0,\infty} \triangleq \max\left\{\left\|\mathrm{x}_0\right\|_0, \left\|\mathrm{x}_1\right\|_0, \ldots, \left\|\mathrm{x}_{(N/a)-1}\right\|_0\right\} \qquad (5)$$

denote the $\ell_{0,\infty}$-norm, which, if applied to the real DGT of an audio signal, measures the maximum number of its non-zero frequency components, where the maximum is taken over time, see also [14].

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

$$\min_{\mathrm{B,y}} \left\|\mathrm{B}\right\|_{0,\infty} \quad \text{s.t.} \quad \mathrm{y} \in \Gamma_{\mathrm{x}} \quad \text{and} \qquad (6a)$$

$$\left\|\eta\left(\mathrm{A_G y}\right) - \mathrm{B}\right\|_{\mathrm{F}} \le \epsilon,$$

$$\min_{\mathrm{B,y}} \left\|\mathrm{B}\right\|_{0,\infty} \quad \text{s.t.} \quad \mathrm{y} \in \Gamma_{\mathrm{x}} \quad \text{and}$$

$$\left\|\mathrm{y} - \mathrm{D_G}\, \eta^{\dagger}\left(\mathrm{B}\right)\right\|_2 \le \epsilon, \qquad (6b)$$

where, again, (6a) and (6b) present the formulation referred to as the analysis and the synthesis variant, respectively.

As in the original SPAIN setting [17], we will adapt the ADMM technique to solve problems (6a) and (6b).

A. A-SPAIN Modified

We will first consider the analysis variant (6a). For fixed sparsity parameter k, define

$$\mathcal{S}_k \triangleq \left\{\mathrm{X} \in \mathbb{C}^{M'\times(N/a)} : \left\|\mathrm{X}\right\|_{0,\infty} \le k\right\}.$$

In order to apply the ADMM algorithm, we rewrite (6a) with $\epsilon = 0$ as

$$\min_{\mathrm{B,y}} \chi_{\mathcal{S}_k}(\mathrm{B}) + \chi_{\Gamma_{\mathrm{x}}}(\mathrm{y}) \quad \text{s.t.} \quad \mathrm{A_G y} - \eta^{\dagger}\left(\mathrm{B}\right) = 0$$

The Augmented Lagrangian is given as,

$$\mathcal{L}_{\delta}(\mathrm{y}, \lambda, \mathrm{B}) = \chi_{\mathcal{S}_k}(\mathrm{B}) + \chi_{\Gamma_{\mathrm{x}}}(\mathrm{y}) + \lambda^{\mathsf{T}}\left(\mathrm{A_G y} - \eta^{\dagger}\left(\mathrm{B}\right)\right) + \frac{\delta}{2}\left\|\mathrm{A_G y} - \eta^{\dagger}\left(\mathrm{B}\right)\right\|_2^2,$$

leading to the Augmented Lagrangian in *scaled form* [55],

$$\mathcal{L}_{\delta}(\mathrm{y}, \mathrm{r}, \mathrm{B}) = \chi_{\mathcal{S}_k}(\mathrm{B}) + \chi_{\Gamma_{\mathrm{x}}}(\mathrm{y}) + \frac{\delta}{2}\left\|\mathrm{A_G y} - \eta^{\dagger}\left(\mathrm{B}\right) + \mathrm{r}\right\|_2^2 - \frac{\delta}{2}\|\mathrm{r}\|_2^2.$$

The update rules of ADMM now yield,

$$\mathrm{B}^{(i+1)} = \arg\min\nolimits_{\mathrm{B}\in\mathcal{S}_k} \left\|\mathrm{A_G y}^{(i)} - \eta^{\dagger}\left(\mathrm{B}\right) + \mathrm{r}^{(i)}\right\|_2 \qquad (7a)$$

$$\mathrm{y}^{(i+1)} = \arg\min\nolimits_{\mathrm{y}\in\Gamma_{\mathrm{x}}} \left\|\mathrm{A_G y} - \eta^{\dagger}\left(\mathrm{B}^{(i+1)}\right) + \mathrm{r}^{(i)}\right\|_2 \qquad (7b)$$

$$\mathrm{r}^{(i+1)} = \mathrm{r}^{(i)} + \mathrm{A_G y}^{(i+1)} - \eta^{\dagger}\left(\mathrm{B}^{(i+1)}\right). \qquad (7c)$$

It is important to observe 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 (7a) is solved exactly [15, p. 42] using the time-frequency hard-thresholding operator $\mathcal{H}_k^{\mathrm{TF}} : \mathbb{C}^{M'\times(N/a)} \to \mathbb{C}^{M'\times(N/a)}$ defined as

$$\mathcal{H}_k^{\mathrm{TF}}\left(\left[\mathrm{x}_0\, \mathrm{x}_1 \cdots \mathrm{x}_{(N/a)-1}\right]\right) = \left[\mathcal{H}_k(\mathrm{x}_0)\, \mathcal{H}_k(\mathrm{x}_1) \cdots \mathcal{H}_k(\mathrm{x}_{(N/a)-1})\right], \qquad (8)$$

where $\mathcal{H}_k(\cdot)$ is—up to pre-scaling[6]—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 (7a) is given by

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

As it is shown in Appendix A, the y-update (7b) can be equivalently computed by an application of the inverse DGT using the canonical dual window to

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

followed by a projection of the outcome onto the set of feasible solutions Γ_{x}.

Remark 1: We note that an analogous statement for the special case of a Parseval frame was shown in [55]. Our proof technique generalizes this result to arbitrary frames satisfying the painless condition (2). Moreover, it takes into account that Γ_{x} as defined in (1) is a set of *real-valued* signals (in contrast to [55]), which makes the situation slightly more complicated. This is caused by the fact that Γ_{x} is *not* an affine subspace over the complex field and some existing results cannot be applied.

The overall algorithm is initialized with a certain pre-defined sparsity parameter $k = s$, which will be augmented in every t^{th} iteration by s, until the stopping criterion is met. Its pseudocode is presented in Alg. 1, where we have used a more convenient matrix notation, i.e., $\mathrm{R}^{(i)} = \eta(\mathrm{r}^{(i)})$. Note that usually $s = t = 1$; however, motivated by the original SPAIN algorithm, we allow for more versatile settings in order to speed up the algorithm ($s > 1$) or promote its convergence ($t > 1$). The input object $\mathrm{D_G^{cd}}$ denotes the synthesis operator of the canonical dual Gabor frame [35]. In practice, of course, only the windows for $\mathrm{A_G}$ and $\mathrm{D_G^{cd}}$ as well as the parameters a and M are passed to the algorithm, and not the full matrices.

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

Algorithm 1: A-SPAIN-MOD

```
   Input: $\mathrm{A_G}$, $\mathrm{D_G^{cd}}$, $\mathrm{M_R x}$, $\Gamma_\mathrm{x}$, $s$, $t$, $\epsilon$
   Output: $\hat{\mathrm{x}}$
1  $\mathrm{y}^{(0)} = \mathrm{M_R x}$, $\mathrm{R}^{(0)} = 0$, $i = 0$, $k = s$
2  $\mathrm{B}^{(i+1)} = \mathcal{H}_k^{\mathrm{TF}}\Big(\eta\big(\mathrm{A_G y}^{(i)}\big) + \mathrm{R}^{(i)}\Big)$
3  $\mathrm{y}^{(i+1)} =$
       $\operatorname*{arg\,min}_{\mathrm{y}\in\Gamma_\mathrm{x}} \Big\| \mathrm{y} - \mathrm{D_G^{cd}}\Big(\eta^\dagger\big(\mathrm{B}^{(i+1)} - \mathrm{R}^{(i)}\big)\Big)\Big\|_2$
4  if $\big\|\eta\big(\mathrm{A_G y}^{(i+1)}\big) - \mathrm{B}^{(i+1)}\big\|_\mathrm{F} \le \epsilon$ then
5      terminate
6  else
7      $\mathrm{R}^{(i+1)} = \mathrm{R}^{(i)} + \eta\big(\mathrm{A_G y}^{(i+1)}\big) - \mathrm{B}^{(i+1)}$
8      $i \leftarrow i + 1$
9      if $i \mod t = 0$ then
10         $k \leftarrow k + s$
11     go to 2
12 return $\hat{\mathrm{x}} = \mathrm{y}^{(i+1)}$
```

Algorithm 2: S-SPAIN-MOD

```
   Input: $\mathrm{D_G}$, $\mathrm{A_G^{cd}}$, $\mathrm{M_R x}$, $\Gamma_\mathrm{x}$, $s$, $t$, $\epsilon$
   Output: $\hat{\mathrm{x}}$
1  $\mathrm{y}^{(0)} = \mathrm{M_R x}$, $\mathrm{r}^{(0)} = 0$, $i = 0$, $k = s$
2  $\mathrm{B}^{(i+1)} = \mathcal{H}_k^{\mathrm{TF}}\Big(\eta\big(\mathrm{A_G^{cd}}\big(\mathrm{y}^{(i)} - \mathrm{r}^{(i)}\big)\big)\Big)$
3  $\mathrm{y}^{(i+1)} =$
       $\operatorname*{arg\,min}_{\mathrm{y}\in\Gamma_\mathrm{x}} \Big\| \mathrm{D_G}\,\eta^\dagger\big(\mathrm{B}^{(i+1)}\big) - \mathrm{y} + \mathrm{r}^{(i)}\Big\|_2$
4  if $\big\|\mathrm{D_G}\,\eta^\dagger\big(\mathrm{B}^{(i+1)}\big) - \mathrm{y}^{(i+1)}\big\|_2 \le \epsilon$ then
5      terminate
6  else
7      $\mathrm{r}^{(i+1)} = \mathrm{r}^{(i)} + \mathrm{D_G}\,\eta^\dagger\big(\mathrm{B}^{(i+1)}\big) - \mathrm{y}^{(i+1)}$
8      $i \leftarrow i + 1$
9      if $i \mod t = 0$ then
10         $k \leftarrow k + s$
11     go to 2
12 return $\hat{\mathrm{x}} = \mathrm{y}^{(i+1)}$
```

B. S-SPAIN Modified

We now consider the synthesis variant (6b). Let us fix a sparsity parameter k and rewrite (6b) with $\epsilon = 0$ as

$$\min_{\mathrm{B},\mathrm{y}} \chi_{\mathcal{S}_k}(\mathrm{B}) + \chi_{\Gamma_\mathrm{x}}(\mathrm{y}) \quad \text{s.t.} \quad \mathrm{D_G}\,\eta^\dagger(\mathrm{B}) - \mathrm{y} = 0$$

The Augmented Lagrangian is given as,

$$\mathcal{L}_\delta(\mathrm{y}, \lambda, \mathrm{B}) = \chi_{\mathcal{S}_k}(\mathrm{B}) + \chi_{\Gamma_\mathrm{x}}(\mathrm{y}) + \lambda^\mathsf{T}\left(\mathrm{D_G}\,\eta^\dagger(\mathrm{B}) - \mathrm{y}\right) + \frac{\delta}{2}\left\|\mathrm{D_G}\,\eta^\dagger(\mathrm{B}) - \mathrm{y}\right\|_2^2,$$

leading to the Augmented Lagrangian in *scaled form* [55],

$$\mathcal{L}_\delta(\mathrm{y}, \mathrm{r}, \mathrm{B}) = \chi_{\mathcal{S}_k}(\mathrm{B}) + \chi_{\Gamma_\mathrm{x}}(\mathrm{y}) + \frac{\delta}{2}\left\|\mathrm{D_G}\,\eta^\dagger(\mathrm{B}) - \mathrm{y} + \mathrm{r}\right\|_2^2 - \frac{\delta}{2}\|\mathrm{r}\|_2^2.$$

The update rules of ADMM now yield,

$$\mathrm{B}^{(i+1)} = \text{arg min}_{\mathrm{B}\in\mathcal{S}_k} \left\|\mathrm{D_G}\,\eta^\dagger(\mathrm{B}) - \mathrm{y}^{(i)} + \mathrm{r}^{(i)}\right\|_2 \tag{9a}$$

$$\mathrm{y}^{(i+1)} = \text{arg min}_{\mathrm{y}\in\Gamma_\mathrm{x}} \left\|\mathrm{D_G}\,\eta^\dagger\left(\mathrm{B}^{(i+1)}\right) - \mathrm{y} + \mathrm{r}^{(i)}\right\|_2 \tag{9b}$$

$$\mathrm{r}^{(i+1)} = \mathrm{r}^{(i)} + \mathrm{D_G}\,\eta^\dagger\left(\mathrm{B}^{(i+1)}\right) - \mathrm{y}^{(i+1)}. \tag{9c}$$

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

The B-update (9a) is a challenging task to solve; in fact, it is a sparse reconstruction problem utilizing the $\ell_{0,\infty}$-norm. However, we will rely on the convenient ADMM behavior that it still converges even when the individual steps are computed only approximately. Adapting the idea in [55] to our setting, we suggest to apply the time-frequency hard-thresholding operator $\mathcal{H}_k^{\mathrm{TF}}$ to the analysis coefficients of $\mathrm{y}^{(i)} - \mathrm{r}^{(i)}$ with respect to the canonical dual frame, or, more formally,

$$\mathrm{B}^{(i+1)} \approx \mathrm{B}_{\text{appr}}^{(i+1)} = \mathcal{H}_k^{\mathrm{TF}}\left(\eta\left(\mathrm{A_G^{cd}}\left(\mathrm{y}^{(i)} - \mathrm{r}^{(i)}\right)\right)\right).$$

Here, $\mathrm{A_G^{cd}}$ denotes the analysis operator of the canonical dual Gabor frame. A formal justification for this approximation is given in Appendix B, extending the argument in [55] to arbitrary frames. Again, the overall algorithm is initialized with a certain pre-defined sparsity parameter $k = s$, which will be augmented in every t^{th} iteration by s, until the stopping criterion is met, see Alg. 2.

Remark 2: Both analysis and synthesis variants, i.e., A-SPAIN-MOD and S-SPAIN-MOD, aim at solving the minimization problem (6) with respect to the $\ell_{0,\infty}$-norm defined in (5). Roughly speaking, A-SPAIN-MOD searches for a signal in the feasible set Γ_x such that its real DGT is maximally sparse in frequency direction *for all* time instants. Similarly, S-SPAIN-MOD searches for TF coefficients which are maximally sparse in frequency direction *for all* time instants such that a signal in the feasible set Γ_x can be synthesized via the inverse real DGT. As a matter of fact, a sparse representation with respect to frequency prohibits the occurrence of peaks, which are typically visible in the real DGT of the gapped signal around the gap borders, see Fig. 1(c). This also illustrates the weakness of "global" ℓ_0-minimization according to (3), i.e., that it does not sufficiently penalize the signal in these specific "local" areas. With $\ell_{0,\infty}$-minimization, 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 the $\ell_{0,\infty}$-approach only exploits sparsity in frequency direction. However, in contrast to audio declipping, the benefit of sparsity in time-direction seems to be limited here.

From an algorithmic perspective we would like to point out that the ADMM algorithm provides an elegant approach to perform (approximate) $\ell_{0,\infty}$-minimization. Optimization with

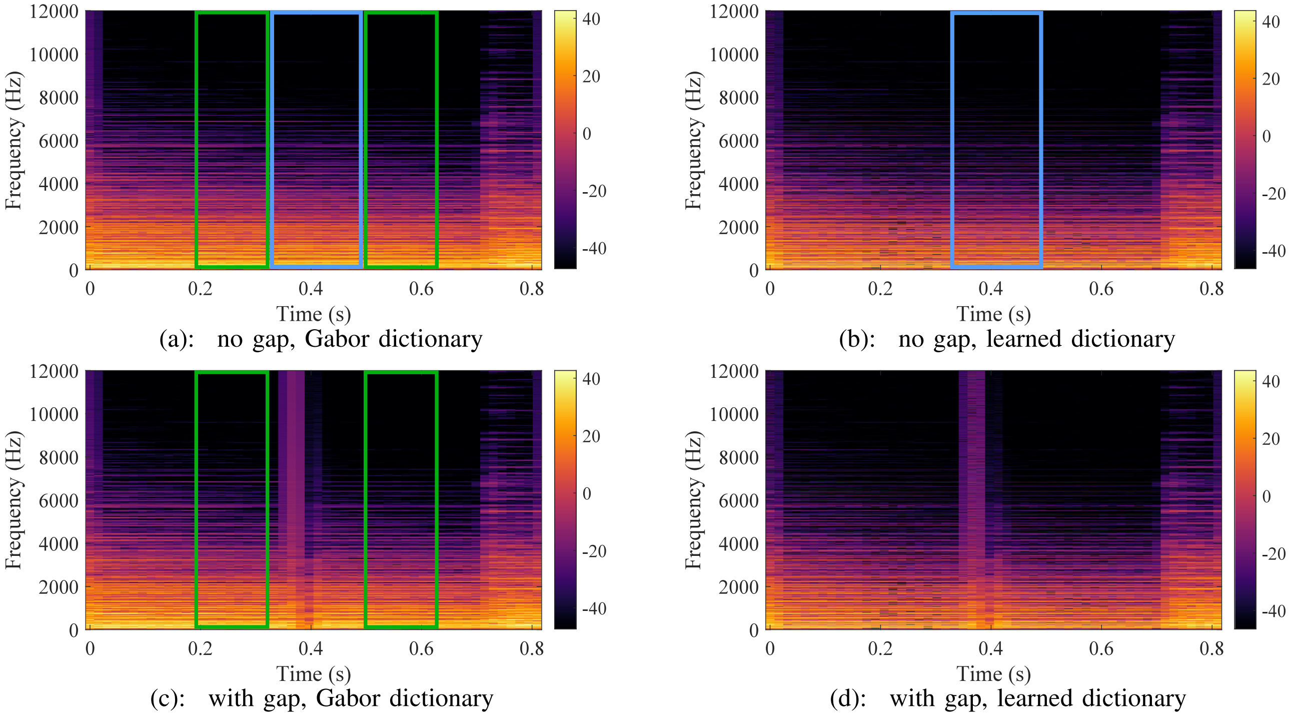


Fig. 1. Analysis coefficients of signal "a60_piano_schubert." (a) and (b) depict the signal without gap, (c) and (d) with gap of length 40 ms. (a) and (c) use Gabor dictionary, (b) and (d) use learned dictionary. The coefficients within the green rectangles are used for training. The sparsity of the coefficients within the blue rectangles is analyzed in Fig. 2.

respect to the mixed $\ell_{0,\infty}$-norm is usually quite involved—not only because of the ℓ_0-component, also the ℓ_∞-component causes difficulties. We will see this in the next section, where a conventional convex relaxation strategy is pursued to learn a dictionary with respect to the $\ell_{0,\infty}$-criterion. In order to make the resulting algorithm tractable in practice, we have to resort to suboptimal simplifications and have to take additional measures to obtain a reasonable substitution of the ℓ_∞-component. This is avoided by the ADMM algorithm used in SPAIN-MOD: $\ell_{0,\infty}$-minimization instead of ℓ_0-minimization is easily implemented by replacing the conventional hard-thresholding operator by the time-frequency hard-thresholding operator $\mathcal{H}_k^{\mathrm{TF}}(\cdot)$ introduced in (8).

V. Dictionary Learning

As mentioned above, the success of all SPAIN variants (including both original and modified versions) heavily depends on the fact that most real-world audio signals have approximately sparse representations with respect to Gabor dictionaries. One could argue, that other dictionaries might allow for even more sparse signal representations and, in turn, enhance the audio inpainting performance, if they are used instead. However, it is by far not immediate how to choose or design a dictionary which exhibits superior audio inpainting capabilities compared with the Gabor dictionary.

The idea, we intend to pursue, is to *learn* an optimized dictionary from the reliable signal parts around the gap with the purpose to obtain also a representation with increased sparsity within the gap and, consequently, improved inpainting quality. Clearly, this relies on the hypothesis that the optimum sparsifying dictionary does not vary too fast. Furthermore, some additional reliable signal parts in the neighborhood of the gap are needed for learning, so that a certain minimum distance between adjacent gaps is desirable, in case more than one gap occurs. In order to keep the learning effort low, we avoid to learn the dictionary from scratch; our approach is to "deform" a given Gabor dictionary.

A. Dictionary Learning Framework

Let us reconsider the modified SPAIN algorithm as discussed in the previous section. According to (6), our goal is to construct 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. Mathematically, we aim at solving

$$\min_{\mathrm{A}} \left\| \left(\eta(\mathrm{Ax})\right)_{\mathcal{N}} \right\|_{0,\infty} \quad \text{s.t.} \quad \mathrm{A}^{\mathsf{H}}\mathrm{A} \text{ is diagonal}, \tag{10}$$

where $\mathcal{N}$ represents the neighborhood. Assume that we are given a Gabor frame satisfying the painless condition (2) with analysis operator A_{G}. Our next step is to "deform" this Gabor frame using a unitary "deformation" operator $\mathrm{W} : \mathbb{C}^P \to \mathbb{C}^P$ by the definition of a modified analysis operator

$$\mathrm{A} = \mathrm{W}\mathrm{A}_{\mathrm{G}}. \tag{11}$$

Here, we restrict to unitary deformations, since these preserve the painless condition (2) due to $\mathrm{A}^{\mathsf{H}}\mathrm{A} = \mathrm{A}_{\mathrm{G}}^{\mathsf{H}}\mathrm{W}^{\mathsf{H}}\mathrm{W}\mathrm{A}_{\mathrm{G}} = \mathrm{A}_{\mathrm{G}}^{\mathsf{H}}\mathrm{A}_{\mathrm{G}} = \mathrm{S}_{\mathrm{G}}$. Moreover, we only care about deformation operators which increase sparsity in frequency direction and which preserve the conjugate-symmetry of the DGT of real-valued signals. Hence, we additionally impose the following structural constraint[7] on W,

$$\mathrm{W} : \mathbb{C}^P \to \mathbb{C}^P, \quad \mathrm{z} \mapsto \tau^{-1}\left(\mathrm{V}\,\tau(\mathrm{z})\right) \tag{12}$$

with[8] $\mathrm{V} \in \mathbb{C}^{M\times M}$ being a unitary matrix with the special form described below. It is evident that (12) represents a well-defined unitary operator. In order to describe the structure of V, we have to distinguish between M even and M odd. For M even, let

$$\mathrm{V} = \mathrm{V}_{\mathrm{e}} = \begin{bmatrix} 1 & & & 0 \\ & \mathrm{U} & & \\ & & 1 & \\ 0 & & & \mathrm{FU^{*}F} \end{bmatrix}, \tag{13}$$

whereas for M odd,

$$\mathrm{V} = \mathrm{V}_{\mathrm{o}} = \begin{bmatrix} 1 & & 0 \\ & \mathrm{U} & \\ 0 & & \mathrm{FU^{*}F} \end{bmatrix} \tag{14}$$

with unitary $\mathrm{U} \in \mathbb{C}^{\lfloor (M-1)/2\rfloor \times \lfloor (M-1)/2\rfloor}$ and flipping matrix

$$\mathrm{F} = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix} \in \mathbb{C}^{\lfloor (M-1)/2\rfloor \times \lfloor (M-1)/2\rfloor}.$$

It is not difficult to see that for a conjugate-symmetric z we can always choose the following representation,

$$\mathrm{W}(\mathrm{z}) = \eta^{\dagger}\left(\mathrm{U}_{\mathrm{e/o}}\,\eta(\mathrm{z})\right) \tag{15}$$

with

$$\mathrm{U}_{\mathrm{e}} = \begin{bmatrix} 1 & & 0 \\ & \mathrm{U} & \\ 0 & & 1 \end{bmatrix} \quad \text{and} \quad \mathrm{U}_{\mathrm{o}} = \begin{bmatrix} 1 & 0 \\ 0 & \mathrm{U} \end{bmatrix}, \tag{16}$$

respectively. Combining (11), (15), and (10), we obtain

$$\min_{\mathrm{U}_{\mathrm{e/o}} \in \mathcal{U}_{\mathrm{e/o}}} \left\| \mathrm{U}_{\mathrm{e/o}}\left(\eta(\mathrm{A}_{\mathrm{G}}\mathrm{x})\right)_{\mathcal{N}} \right\|_{0,\infty},$$

where $\mathcal{U}_{\mathrm{e/o}}$ denotes the set of all unitary matrices $\mathrm{U}_{\mathrm{e/o}}$ with the structure described in (16). Observe, however, that this is a highly non-convex problem, so that further measures have to be taken in order to obtain a solution.

To this end, we pursue the classical approach to *relax* ℓ_0-norms to ℓ_1-norms [15], [45]. 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[9]). This yields the following minimization problem,

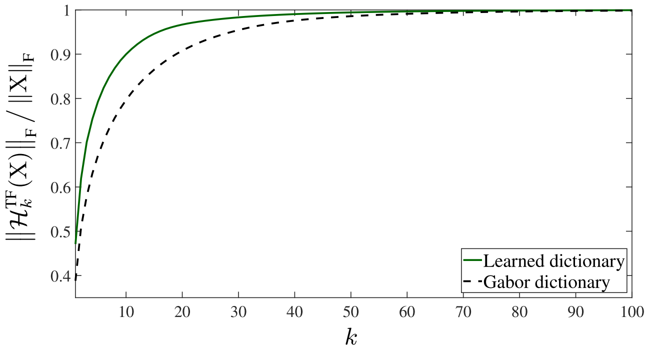


Fig. 2. Sparsity of analysis coefficients of signal “a60_piano_schubert,” using the time-frequency hard-thresholding operator $\mathcal{H}_k^{\mathrm{TF}}$ applied to the coefficients X within the blue rectangle of Fig. 1(a) for the Gabor dictionary and the blue rectangle of Fig. 1(b) for the learned dictionary, respectively.

$$\hat{\mathrm{U}}_{\mathrm{e/o}} = \arg\min_{\mathrm{U}_{\mathrm{e/o}} \in \mathcal{U}_{\mathrm{e/o}}} \sum_{q\in\mathcal{N}} \left\| \mathrm{U}_{\mathrm{e/o}}\left(\eta(\mathrm{A}_{\mathrm{G}}\mathrm{x})\right)_q \right\|_1. \tag{17}$$

Remark 3: We emphasize that the second replacement, i.e., $\max$ to summation, could be omitted. Apparently, retaining the $\max$-operation would be desirable from the conceptual point of view, we have stressed above. But in our implementations the summation variant was significantly faster than the $\max$ variant. In fact, the replacement even turned out to be mandatory in order to avoid exceeding the computational resources of our simulation framework. Note that the summation variant optimizes the deformation matrix W such that the sparsity with respect to frequency is maximized[10] *on average* for all time instants in the considered neighborhood $\mathcal{N}$. We are aware that this approach does not guarantee that the sparsity is increased equally for all time instants in $\mathcal{N}$ but for time instants within and around the gap (used for inpainting) we can still expect a considerably improved sparsity of the original signal. We further facilitate this by restrictions via ρ_{start}, d, and $r_{\max}$ defined in Subsection V-B, which 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. Our numerical experiments also confirm this behavior, see Subsection VI-A, and in particular Fig. 2, which depicts the sparsity gain achieved by the learned dictionary in terms of the *time-frequency* hard-thresholding operator $\mathcal{H}_k^{\mathrm{TF}}$, i.e., using an $\ell_{0,\infty}$-measure. Hence, although suboptimal, the chosen average optimality criterion seems to be reasonable for the learning phase.

Finally, observe that the optimization (17) is effectively carried out over the set $\mathcal{U}$ of unitary $\lfloor (M-1)/2\rfloor \times \lfloor (M-1)/2\rfloor$ matrices U according to (16) with optimum $\hat{\mathrm{U}}$ given by

$$\hat{\mathrm{U}} = \arg\min\nolimits_{\mathrm{U}\in\mathcal{U}} \sum_{q\in\mathcal{N}} \left\| \mathrm{U}\mathrm{E}_{\mathrm{e/o}}\left(\eta(\mathrm{A}_{\mathrm{G}}\mathrm{x})\right)_q \right\|_1, \tag{18}$$

where

[7] Roughly speaking, this construction avoids the mixing of “positive” and “negative” frequencies.

[8] In order to prevent confusion, we emphasize that $\mathrm{V}\,\tau(\mathrm{z})$ is a conventional matrix-matrix product.

[9] Note that the $\ell_{1,1}$-norm of a matrix is the same as the ℓ_1-norm of the vector consisting of the matrix elements.

[10] To prevent any confusion with this statement: sparsity maximization corresponds to minimization of the number of non-zeros.

```
Algorithm 3: A-SPAIN-LEARNED
   Input: $\mathrm{A_G}$, $\mathrm{D_G^{cd}}$, $\mathrm{U_{e/o}}$, $\mathrm{M_R x}$, $\Gamma_\mathrm{x}$, $s$, $t$, $\epsilon$
   Output: $\hat{\mathrm{x}}$
 1 $\mathrm{y}^{(0)} = \mathrm{M_R x}$, $\mathrm{R}^{(0)} = 0$, $i = 0$, $k = s$
 2 $\mathrm{B}^{(i+1)} = \mathcal{H}_k^{\mathrm{TF}}\Big(\mathrm{U_{e/o}}\eta\big(\mathrm{A_G y}^{(i)}\big) + \mathrm{R}^{(i)}\Big)$
 3 $\mathrm{y}^{(i+1)} = \operatorname*{arg\,min}_{\mathrm{y}\in\Gamma_\mathrm{x}} \Big\| \mathrm{y} - \mathrm{D_G^{cd}}\Big(\eta^\dagger\big(\mathrm{U_{e/o}^H B}^{(i+1)} - \mathrm{U_{e/o}^H R}^{(i)}\big)\Big)\Big\|_2$
 4 if $\big\|\mathrm{U_{e/o}}\eta\big(\mathrm{A_G y}^{(i+1)}\big) - \mathrm{B}^{(i+1)}\big\|_\mathrm{F} \leq \epsilon$ then
 5     terminate
 6 else
 7     $\mathrm{R}^{(i+1)} = \mathrm{R}^{(i)} + \mathrm{U_{e/o}}\eta\big(\mathrm{A_G y}^{(i+1)}\big) - \mathrm{B}^{(i+1)}$
 8     $i \leftarrow i + 1$
 9     if $i \mod t = 0$ then
10         $k \leftarrow k + s$
11     go to 2
12 return $\hat{\mathrm{x}} = \mathrm{y}^{(i+1)}$
```

```
Algorithm 4: S-SPAIN-LEARNED
   Input: $\mathrm{D_G}$, $\mathrm{A_G^{cd}}$, $\mathrm{U_{e/o}}$, $\mathrm{M_R x}$, $\Gamma_\mathrm{x}$, $s$, $t$, $\epsilon$
   Output: $\hat{\mathrm{x}}$
 1 $\mathrm{y}^{(0)} = \mathrm{M_R x}$, $\mathrm{r}^{(0)} = 0$, $i = 0$, $k = s$
 2 $\mathrm{B}^{(i+1)} = \mathcal{H}_k^{\mathrm{TF}}\Big(\mathrm{U_{e/o}}\eta\big(\mathrm{A_G^{cd}}\big(\mathrm{y}^{(i)} - \mathrm{r}^{(i)}\big)\big)\Big)$
 3 $\mathrm{y}^{(i+1)} = \operatorname*{arg\,min}_{\mathrm{y}\in\Gamma_\mathrm{x}} \Big\| \mathrm{D_G}\,\eta^\dagger\big(\mathrm{U_{e/o}^H B}^{(i+1)}\big) - \mathrm{y} + \mathrm{r}^{(i)}\Big\|_2$
 4 if $\big\|\mathrm{D_G}\,\eta^\dagger\big(\mathrm{U_{e/o}^H B}^{(i+1)}\big) - \mathrm{y}^{(i+1)}\big\|_2 \leq \epsilon$ then
 5     terminate
 6 else
 7     $\mathrm{r}^{(i+1)} = \mathrm{r}^{(i)} + \mathrm{D_G}\,\eta^\dagger\big(\mathrm{U_{e/o}^H B}^{(i+1)}\big) - \mathrm{y}^{(i+1)}$
 8     $i \leftarrow i + 1$
 9     if $i \mod t = 0$ then
10         $k \leftarrow k + s$
11     go to 2
12 return $\hat{\mathrm{x}} = \mathrm{y}^{(i+1)}$
```

$$\begin{aligned}\mathrm{E_e} &= \left[0_{\lfloor (M-1)/2\rfloor\times 1} \;\; \mathrm{I}_{\lfloor (M-1)/2\rfloor} \;\; 0_{\lfloor (M-1)/2\rfloor\times 1}\right],\\ \mathrm{E_o} &= \left[0_{\lfloor (M-1)/2\rfloor\times 1} \;\; \mathrm{I}_{\lfloor (M-1)/2\rfloor}\right].\end{aligned}$$

Let us assume we have found an optimized unitary matrix $\hat{\mathrm{U}}$ (we will present an algorithm for this task in the next subsection). Inserting it into (16) and (13)/(14) yields $\hat{\mathrm{U}}_\mathrm{e/o}$ and $\hat{\mathrm{V}}_\mathrm{e/o}$, respectively, and, furthermore, via (12) and (11), we obtain the analysis operator of the deformed frame as,

$$\hat{\mathrm{A}} : \mathbb{C}^N \to \mathbb{C}^P, \quad \mathrm{y} \mapsto \tau^{-1}\left(\hat{\mathrm{V}}_\mathrm{e/o}\,\tau(\mathrm{A_G y})\right). \tag{19}$$

Note that the synthesis operator of its canonical dual frame is given by

$$\hat{\mathrm{D}}^\mathrm{cd} : \mathbb{C}^P \to \mathbb{C}^N, \quad \mathrm{z} \mapsto \mathrm{D_G^{cd}}\tau^{-1}\left(\hat{\mathrm{V}}_\mathrm{e/o}^\mathrm{H}\tau(\mathrm{z})\right). \tag{20}$$

This puts us in the position to reformulate Algs. 1 and 2 in a way such that the learned sparsity-optimized frame is used instead of the Gabor frame. More specifically, for the analysis variant we replace $\mathrm{A_G}$ with $\hat{\mathrm{A}}$ and $\mathrm{D_G^{cd}}$ with $\hat{\mathrm{D}}^\mathrm{cd}$. In the algorithm, the (optimized) matrix $\hat{\mathrm{U}}_\mathrm{e/o}$ is required[11] as an additional input, see Alg. 3.

For the synthesis variant, we suggest to still deform the analysis operator, but here the one of the canonical dual frame. Then, the synthesis operator of the corresponding original frame is expected to represent any signal in the feasible set Γ_x with fewer coefficients than the synthesis operator of the given Gabor frame. Again, the algorithm requires[11] the specification of the (optimized) matrix $\hat{\mathrm{U}}_\mathrm{e/o}$ as additional input, see Alg. 4.

B. Basis Optimization Technique

In order to solve (18), we adapt a basis optimization technique originally developed in the context of channel estimation [56], [57]. To simplify notation we set $\tilde{M} \triangleq \lfloor (M-1)/2 \rfloor$ and define $\eta_\mathrm{e/o}(\cdot) \triangleq \mathrm{E_{e/o}}\eta(\cdot)$ Because the minimization problem (18) is non-convex (since $\mathcal{U}$ is not a convex set), standard convex optimization techniques cannot be used. We therefore propose an approximate iterative algorithm that relies on the following facts [58, p. 8], [59].

- Every unitary $\tilde{M} \times \tilde{M}$ matrix U can be represented in terms of a Hermitian $\tilde{M} \times \tilde{M}$ matrix H as $\mathrm{U} = e^{\mathrm{iH}}$.
- The matrix exponential $\mathrm{U} = e^{\mathrm{iH}}$ can be approximated by its first-order Taylor expansion, i.e., $\mathrm{U} \approx \mathrm{I}_{\tilde{M}} + \mathrm{iH}$.

Even though U is unitary and $\mathrm{I}_{\tilde{M}} + \mathrm{iH}$ is not, this approximation will be close if $\|\mathrm{H}\|_{\infty,\infty}$ is small. Because of this condition, we construct U iteratively: starting with the identity matrix, we perform a *small* update at each iteration, using the approximation $\mathrm{U} \approx \mathrm{I}_{\tilde{M}} + \mathrm{iH}$ in the optimization criterion *but not for actually updating* U (thus, the iterated U is always unitary). More specifically, at the rth iteration, we consider the following update of the unitary matrix $\mathrm{U}^{(r)}$:

$$\mathrm{U}^{(r+1)} = e^{\mathrm{iH}^{(r)}}\mathrm{U}^{(r)},$$

where $\mathrm{H}^{(r)}$ is a small (with respect to $\|\cdot\|_{\infty,\infty}$) Hermitian matrix that remains to be optimized. Note that $\mathrm{U}^{(r+1)}$ is again unitary because both $\mathrm{U}^{(r)}$ and $e^{\mathrm{iH}^{(r)}}$ are unitary.

Ideally, we would like to optimize $\mathrm{H}^{(r)}$ according to (18), i.e., by minimizing

$$\begin{aligned}&\sum_{q\in\mathcal{N}} \left\|\mathrm{U}^{(r+1)}\mathrm{E_{e/o}}\left(\eta(\mathrm{A_G x})\right)_q\right\|_1\\ &\qquad = \sum_{q\in\mathcal{N}} \left\|\mathrm{U}^{(r+1)}\left(\eta_\mathrm{e/o}(\mathrm{A_G x})\right)_q\right\|_1\\ &\qquad = \sum_{q\in\mathcal{N}} \left\|e^{\mathrm{iH}^{(r)}}\mathrm{U}^{(r)}\left(\eta_\mathrm{e/o}(\mathrm{A_G x})\right)_q\right\|_1.\end{aligned}$$

Since this problem is still non-convex, we use the approximation $e^{\mathrm{iH}} \approx \mathrm{I}_{\tilde{M}} + \mathrm{iH}$, and thus the final minimization problem at the

[11] The description of the algorithms with $\hat{\mathrm{U}}_\mathrm{e/o}$ as input parameter is obtained from (19) and (20) via (15), (11), and (12).

rth iteration is

$$\hat{\mathrm{H}}^{(r)} = \arg\min_{\mathrm{H}\in\mathcal{H}_r} \sum_{q\in\mathcal{N}} \left\| (\mathrm{I}_{\tilde{M}} + \mathrm{iH})\, \mathrm{U}^{(r)} \, (\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_q \right\|_1. \tag{21}$$

Here, $\mathcal{H}_r$ is the set of all Hermitian $\tilde{M} \times \tilde{M}$ matrices H that are small in the sense that $\|\mathrm{H}\|_{\infty,\infty} \leq \rho_r$, where ρ_r is a positive constraint level (a small ρ_r ensures a good accuracy of our approximation $\mathrm{U} \approx \mathrm{I}_{\tilde{M}} + \mathrm{iH}$ and also that $e^{\mathrm{i}\hat{\mathrm{H}}^{(r)}}$ is close to $\mathrm{I}_{\tilde{M}}$). The problem (21) is convex and thus can be solved by standard convex optimization techniques [60].

The next step at the rth iteration is to test whether the cost function is smaller for the new unitary matrix $e^{\mathrm{i}\hat{\mathrm{H}}^{(r)}}\mathrm{U}^{(r)}$, i.e., whether

$$\sum_{q\in\mathcal{N}} \left\| e^{\mathrm{i}\hat{\mathrm{H}}^{(r)}}\, \mathrm{U}^{(r)} \, (\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_q \right\|_1 < \sum_{q\in\mathcal{N}} \left\| \mathrm{U}^{(r)} \, (\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_q \right\|_1.$$

In the positive case, we actually perform the update of $\mathrm{U}^{(r)}$ and we retain the constraint level ρ_r for the next iteration:

$$\mathrm{U}^{(r+1)} = e^{\mathrm{i}\hat{\mathrm{H}}^{(r)}}\mathrm{U}^{(r)}, \qquad \rho_{r+1} = \rho_r\,.$$

Otherwise, we reject the update of $\mathrm{U}^{(r)}$ and reduce the constraint level ρ_r:

$$\mathrm{U}^{(r+1)} = \mathrm{U}^{(r)}, \qquad \rho_{r+1} = \frac{\rho_r}{2}\,.$$

By this construction, the cost function sequence $\sum_{q\in\mathcal{N}} \left\| \mathrm{U}^{(r)} (\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_q \right\|_1$, $r = 0, 1, \ldots$ is guaranteed to be monotonically decreasing.

The above iteration process is terminated if ρ_r falls below a prescribed threshold or if the number of iterations exceeds a certain value. The iteration process is initialized by the identity matrix $\mathrm{I}_{\tilde{M}}$, because the "undeformed" Gabor dictionary is known to yield relatively sparse analysis coefficients. We note that efficient algorithms for computing the matrix exponentials $e^{\mathrm{i}\hat{\mathrm{H}}^{(r)}}$ exist [59].

Finally, we would like to emphasize that if we restrict the minimization of (21) to the set $\mathcal{H}_{r,d}$ of small Hermitian matrices with at most d non-vanishing off-diagonals, we obtain a potentially suboptimal variant of the algorithm with reduced computational complexity. In some cases, the choice of such a modification will be mandatory; especially, if we deal with large-scale problems. We note that we used the convex optimization package CVX [61] for minimizing (21) in all our simulations. The overall basis optimization algorithm (including the off-diagonal parameter d as input) is summarized in Alg. 5.

C. Choosing the Learning Neighborhood $\mathcal{N}$

It remains to address the question, how to select the neighborhood $\mathcal{N}$ of the gap, which is needed to specify the (input) training matrix $(\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_{\mathcal{N}}$ for Alg. 5. Apparently, $\mathcal{N}$ should be as close as possible to the gap since this increases the likelihood that the learned dictionary represents the signal part within the gap sparsely. On the other hand, we have to account for small "guard" intervals between the gap and training borders, in order to avoid that unreliable data from the gap influences our training data. Note that the underlying Gabor structure implies that this guard interval should be at least of length w_{g} (length of Gabor window). Regarding the neighborhood size we have to rely on intuition to some extent. A larger neighborhood causes a larger training matrix $(\eta(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_{\mathcal{N}}$ and, probably, more accurate learning results but it also increases chances that signal variations occur within the neighborhood. Such signal variations could yield a dictionary matched to signal parts which are essentially unrelated to the signal within the gap. Suppose the signal $\mathrm{x} \in \mathbb{R}^N$ has a gap at the indices $n = n_{\mathrm{B}}, n_{\mathrm{B}} + 1, \ldots, n_{\mathrm{E}}$. Excluding the guard intervals mentioned above, the neighborhood is given by $\mathcal{N} = \mathcal{N}_{\mathrm{B}} \cup \mathcal{N}_{\mathrm{E}}$ with[12]

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

i.e., a part before and a part after the gap. Its total length is $2L_{\mathcal{N}}$. We will investigate the choice of the length parameter $L_{\mathcal{N}}$ via numerical experiments, see Subsections VI-A and VI-B.

Algorithm 5: Basis Optimization

```
Input: $(\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_{\mathcal{N}}$, $\rho_{\text{start}}$, $d$, $\kappa$, $r_{\max}$
Output: $\hat{\mathrm{U}}$
$\mathrm{U}^{(0)} = \mathrm{I}_{\tilde{M}}$, $\rho_0 = \rho_{\text{start}}$, $r = 0$
$s_0 = \sum_{q\in\mathcal{N}} \left\| \mathrm{U}^{(0)} (\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_q \right\|_1$
while $\rho_r > \kappa$ and $r < r_{\max}$ do
    $\hat{\mathrm{H}}^{(r)} = \arg\min_{\mathrm{H}\in\mathcal{H}_{r,d}} \sum_{q\in\mathcal{N}} \left\| (\mathrm{I}_{\tilde{M}} + \mathrm{iH})\, \mathrm{U}^{(r)} (\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_q \right\|_1$
    $s_{r+1} = \sum_{q\in\mathcal{N}} \left\| e^{\mathrm{i}\hat{\mathrm{H}}^{(r)}}\, \mathrm{U}^{(r)} (\eta_{\text{e/o}}(\mathrm{A}_{\mathrm{G}}\mathrm{x}))_q \right\|_1$
    if $s_{r+1} < s_r$ then
        $\mathrm{U}^{(r+1)} = e^{\mathrm{i}\hat{\mathrm{H}}^{(r)}}\mathrm{U}^{(r)}, \quad \rho_{r+1} = \rho_r$
    else
        $\mathrm{U}^{(r+1)} = \mathrm{U}^{(r)}, \quad \mathrm{s}_{r+1} = \mathrm{s}_r, \quad \rho_{r+1} = \frac{\rho_r}{2}$
    $r \leftarrow r + 1$
return $\hat{\mathrm{U}} = \mathrm{U}^{(r+1)}$
```

D. Discussion

One might ask, whether the proposed dictionary learning framework is only suitable to improve A-SPAIN-MOD and S-SPAIN-MOD or if it can be applied to other sparse inpainting methods such as (re-)weighted ℓ_1-minimization techniques (and generalizations) [18] as well. This is a topic for future research but we believe that appropriate adaptations hold also the potential to yield major performance improvements. Moreover, a closer analysis of the original SPAIN implementation reveals that the window-based segmentation process in combination with the overlap-add technique and the redundant DFT dictionary essentially corresponds to implementations of Gabor analysis and synthesis. Therefore, original SPAIN shares many

[12] Here, it is assumed that the gap is sufficiently centered within $\{0, \ldots, (N/a)-1\}$, so that $0 \leq \lfloor (n_{\mathrm{B}} - w_{\mathrm{g}})/a \rfloor - L_{\mathcal{N}}$ and $\lceil (n_{\mathrm{E}} + w_{\mathrm{g}})/a \rceil + L_{\mathcal{N}} < N/a$.

similarities with SPAIN-MOD although it differs with respect to the application of the ADMM algorithm. Nevertheless, it seems to be rather straightforward to apply our dictionary framework also to the original SPAIN variants. We expect that the main step is to transform the redundant DFT dictionary using the matrix V, see (12), (13), and (14), which does not depend on the actual segment. But note that we are convinced that our suggested SPAIN-MOD should be used in any case (in Subsection VI-D it is shown that SPAIN-MOD outperforms the original SPAIN even without dictionary learning).

VI. Simulation Results

Here, we present a numerical assessment of our dictionary learning technique for sparse audio inpainting. In order to allow for a valid and comparable evaluation, we consider essentially the same setup as in [18]. As main performance criterion, we use the signal-to-distortion ratio (SDR), which is defined as[13]

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

where x_{orig} and x_{inp} denote original and inpainted signal within the gaps, respectively. Obviously, higher SDR values reflect superior reconstruction. As in [18], we compute the average SDR by first calculating the particular values of SDR in dB, and then taking the average. Furthermore, we also compute the PEMO-Q measure [62], which includes a model of the human auditory system. Thus, it is closer to the subjective evaluation than the SDR. The measured quantity is denoted as *objective difference grade* (ODG) and can be interpreted as the degree of perceptual similarity between original and inpainted signal. The ODG attains values from 0 (imperceptible) to -4 (very annoying), thereby, expressing the effect of audio artifacts in the reconstructed signal.

We use a collection of ten music recordings chosen from the EBU SQAM dataset [63] and sampled at 44.1 kHz, with different degrees of sparsity with respect to the original Gabor dictionary. In each test instance, the input was a signal with 5 gaps at random positions. The lengths of these gaps ranged from 5 ms up to 50 ms. For fixed lengths, the results over all ten signals containing the 5 gaps were averaged.

Throughout our experiments we used a tight Gabor frame with the Hann window of length $w_{\mathrm{g}} = 2800$ samples (approximately 64 ms), hop size $a = 700$ samples, and with $M = 2800$ modulations. We used the fast implementation of Gabor transforms provided by the LTFAT toolbox [53], [54] taking into account its time-frequency conventions. In order to keep the computational complexity of the basis optimization algorithm low, we set its maximum number of iterations to $r_{\max} = 20$, its off-diagonal parameter to $d = 1$ (except where noted otherwise), and its remaining parameters to $\rho_{\text{start}} = 1$ and $\kappa = 2^{-20}$. Finally, all SPAIN variants used the input parameters $s = t = 1$ and $\epsilon = 0.001$.

A. Sparsity With Respect to Learned Dictionary

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

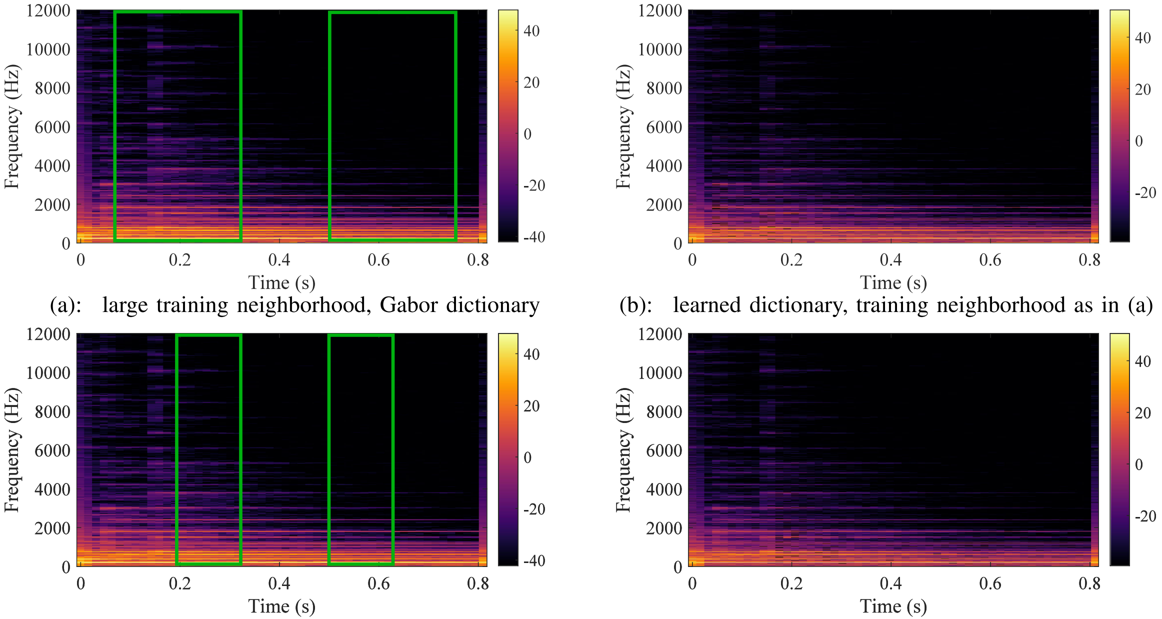
Fig. 1 illustrates the analysis coefficients of the signal "a60_piano_schubert" from the EBU SQAM dataset [63]. Subfigures (a) and (b) depict the signal without gap, (c) and (d) with gap of length 40 ms. Subfigures (a) and (c) use the Gabor dictionary, (b) and (d) the learned dictionary. The coefficients within the green rectangles are used for training, corresponding to a neighborhood length parameter of $L_{\mathcal{N}} = 2w_{\mathrm{g}}/a = 8$. As expected, the learned dictionary allows for more sparse representations than the Gabor dictionary. In particular, the learned dictionary also sparsifies the original signal within the gap area even though the training coefficients are a certain distance apart (from the gap area). This is also confirmed by a quantitative analysis using the time-frequency hard-thresholding operator $\mathcal{H}_k^{\mathrm{TF}}$ applied to the coefficients X within the blue rectangles of Fig. 1(a) and 1(b), which are relevant for the inpainting performance. More specifically, Fig. 2 depicts the normalized Froebenius norm of the k largest coefficients in frequency direction for each time instant specified by X, i.e., $\|\mathcal{H}_k^{\mathrm{TF}}(\mathrm{X})\|_{\mathrm{F}}/\|\mathrm{X}\|_{\mathrm{F}}$, for both Gabor and learned dictionary. It is seen that the same number of k coefficients in frequency direction contain significantly more energy for the learned dictionary than for the Gabor dictionary, or, conversely, fewer coefficients are needed to represent the signal with prescribed accuracy for the learned dictionary than for the Gabor dictionary. Since X is chosen according to the blue rectangles in Fig. 1(a) and 1(b) (and, hence, does not intersect the learning area framed in green), our assumption that the sparsifying dictionary does not change too fast over time is confirmed.

Moreover, we would like to develop an understanding how the size of the learning neighborhood $\mathcal{N}$ around the gap impacts the sparsity of the learned representation. As an example, Fig. 3 depicts the analysis coefficients of the signal "a25_harp" from the EBU SQAM dataset [63]. Subfigures (a) and (c) use the 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 $L_{\mathcal{N}} = 4w_{\mathrm{g}}/a = 16$, whereas the neighborhood length parameter used to obtain (d) was $L_{\mathcal{N}} = 2w_{\mathrm{g}}/a = 8$. Looking very carefully it seems that the analysis coefficients with respect to the dictionary obtained by a larger training set are slightly more sparse. However, especially within the relevant gap area, it is almost impossible to notice any differences, so that a more informative comparison has to be done in terms of inpainting performance, see the next subsection.

B. Performance Comparison: Training Neighborhood Size and Learning Complexity

Here, we evaluate how the size of the training neighborhood impacts the inpainting performance. To that end, we

[13] Often also denoted as signal-to-noise ratio, see, e.g., [1], [18].



(a): large training neighborhood, Gabor dictionary

(b): learned dictionary, training neighborhood as in (a)

(c): medium training neighborhood, Gabor dictionary

(d): learned dictionary, training neighborhood as in (c)

Fig. 3. Analysis coefficients of signal “a25_harp” without gap. (a) and (c) use Gabor dictionary, (b) and (d) use learned dictionaries 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).

considered the same two cases as in the previous subsection, i.e., the two different neighborhoods with length parameters $L_{\mathcal{N}} = 4w_{\mathrm{g}}/a = 16$ and $L_{\mathcal{N}} = 2w_{\mathrm{g}}/a = 8$ and, additionally, a small neighborhood with $L_{\mathcal{N}} = w_{\mathrm{g}}/a = 4$. Furthermore, we study if additional gains can be achieved by increasing the learning complexity via the off-diagonal parameter d. Fig. 4 depicts the inpainting results after averaging over all ten signals with five gaps for the analysis variant A-SPAIN-LEARNED and the synthesis variant S-SPAIN-LEARNED using off-diagonal parameters $d = 1, 3, 5$.

It is seen that the larger off-diagonal parameters $d = 3, 5$ corresponding to higher computational complexity do not yield any improvements. Moreover, $L_{\mathcal{N}} = 8$ is superior to $L_{\mathcal{N}} = 16$ and $L_{\mathcal{N}} = 4$ for the majority of gap lengths in terms of SDR and ODG and this behavior is observed for both analysis and synthesis variants. It seems that the medium size neighborhood is a good compromise between benefits of only using learning coefficients close to the gap and potential disadvantages of small training sets. It is quite remarkable that a larger off-diagonal parameter ($d = 3, 5$) even deteriorates the performance although one would expect increased sparsity. On the other hand, the smaller the off-diagonal parameter the smaller the number of Gabor coefficients which are linearly combined (in frequency direction) to obtain a learned dictionary coefficient. This implies that the choice $d = 1$ will not sparsify peaks in frequency direction to the same extent as it will sparsify the original signal (which is used for training), see Fig. 1(d), so that the chance that SPAIN selects an incorrect signal from the feasible set Γ_{x} is reduced.

Based on these insights, we restricted ourselves to the training neighborhood with length parameter $L_{\mathcal{N}} = 2w_{\mathrm{g}}/a = 8$ and the off-diagonal parameter $d = 1$ for all further comparisons.

C. Selected Time-Domain Results

Next, we show some example signals obtained by our inpainting methods. To that end, Fig. 5 depicts the time-domain signals computed by A-SPAIN-LEARNED and S-SPAIN-LEARNED to fill a 40 ms gap in the signal “a25_harp.” The original signal is included for reference. It is seen that both solutions approximate the original signal quite accurately. In Fig. 6, we plot the graphs of the learned dictionary atoms in the time domain. In particular, we show the atoms corresponding to the first 5 frequency channels obtained by Alg. 5 when applied to the signal “a25_harp” with a 40 ms gap. For comparison, Fig. 6 also contains the Gabor atoms corresponding to the first 5 frequency channels. We can clearly observe differences between the Gabor and learned dictionary atoms but also shared similarities.

D. Performance Comparison: Different Methods

Finally, we will compare our proposed methods with various other existing audio inpainting methods. We largely adopted the simulation settings considered in [18] in order to provide a consistent evaluation. At this point we would like to thank the authors of [18] for their reproducible research policy, which eased this task significantly. Besides our methods [A-SPAIN-MOD, S-SPAIN-MOD, A-SPAIN-LEARNED, S-SPAIN-LEARNED], we used the following audio inpainting techniques (the abbreviations in the rectangular brackets are used in Fig. 7):

- The original SPAIN algorithm introduced in [17] using a frame-wise DFT dictionary with redundancy 4. We considered both analysis variant [A-SPAIN] and synthesis variant with hard thresholding [S-SPAIN H].

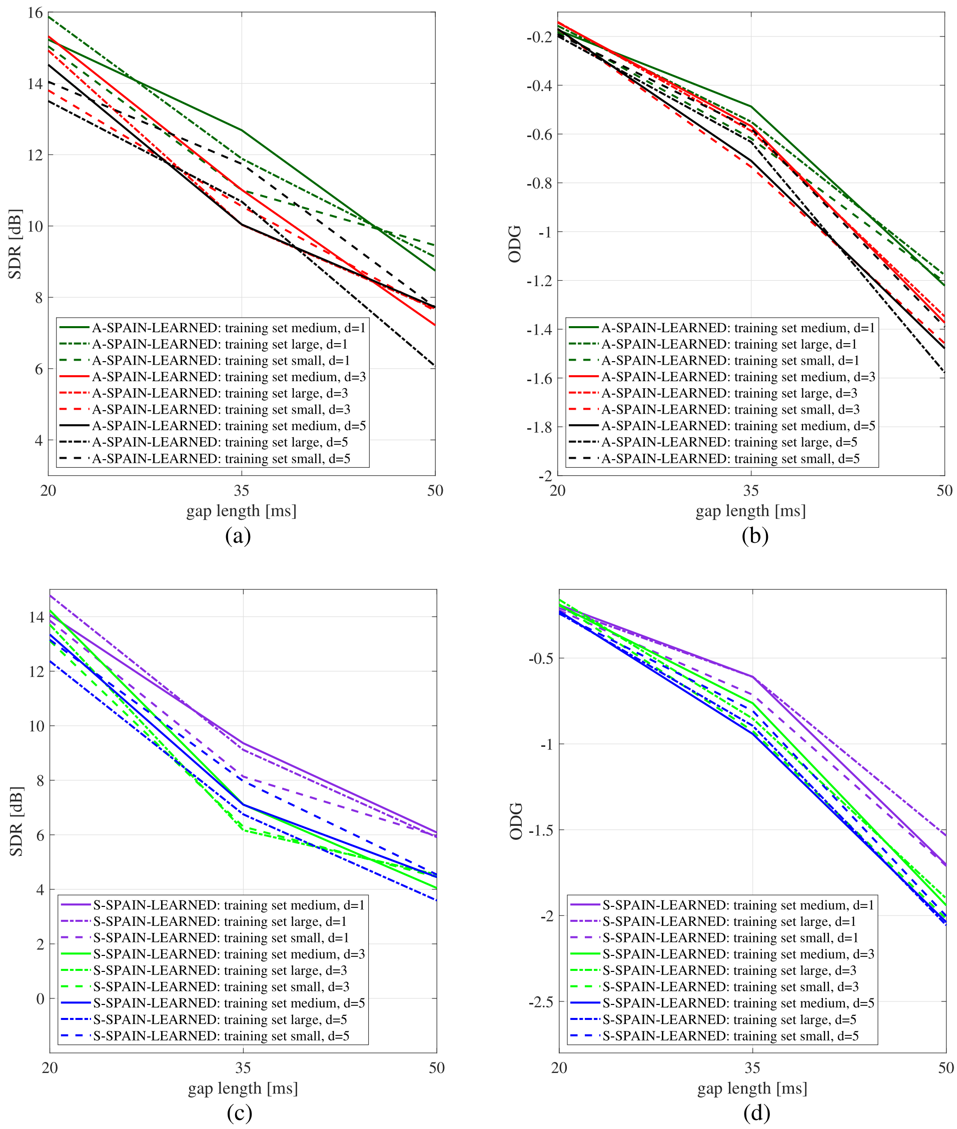


Fig. 4. Comparison of audio inpainting performance using A-SPAIN-LEARNED and S-SPAIN-LEARNED for three different sizes of the training neighborhood and for off-diagonal parameters $d = 1, 3, 5$. (a) and (c) depict results in terms of SDR, (b) and (d) in terms of ODG values.

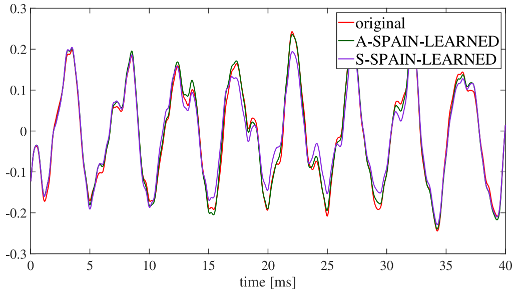


Fig. 5. Audio inpainting results obtained by A-SPAIN-LEARNED and S-SPAIN-LEARNED for signal "a25_harp" with gap of length 40 ms. Original time-domain signal as well as inpainted signals are shown within gap range.

- The frame-wise Janssen algorithm [5] with autoregressive model order $p = \min(3H + 2, w_g/3)$, where H denotes the number of missing/unreliable samples within the current frame (window), and the number of iterations was set to 50 [JANSSEN].
- The weighted Douglas-Rachford algorithm [18], i.e., synthesis model ℓ_1-minimization with ℓ_2-norm-based weighting [DR].
- The weighted Chambolle-Pock algorithm [18], i.e., analysis model ℓ_1-minimization with ℓ_2-norm-based weighting [CP].
- The weighted Chambolle-Pock algorithm, i.e., analysis model ℓ_1-minimization with energy-based weighting, including time-domain compensation [18] for energy loss (number of artificial gaps: 4, number of segments: 10, segment length: quarter of gap length, shift parameter: $w_g/2$) [TDC].

Fig. 7 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, S-SPAIN-MOD, A-SPAIN-LEARNED, and

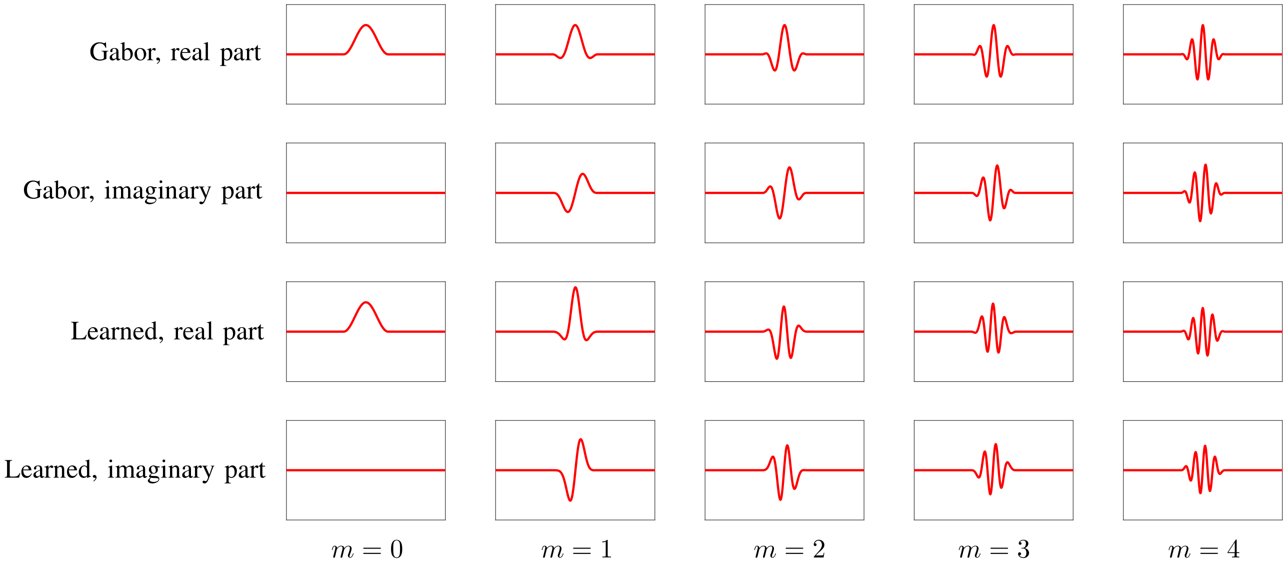


Fig. 6. Gabor atoms versus learned dictionary atoms. Real and imaginary parts of atoms corresponding to frequency channels $m = 0, \ldots, 4$ are shown. All atom plots are equally normalized.

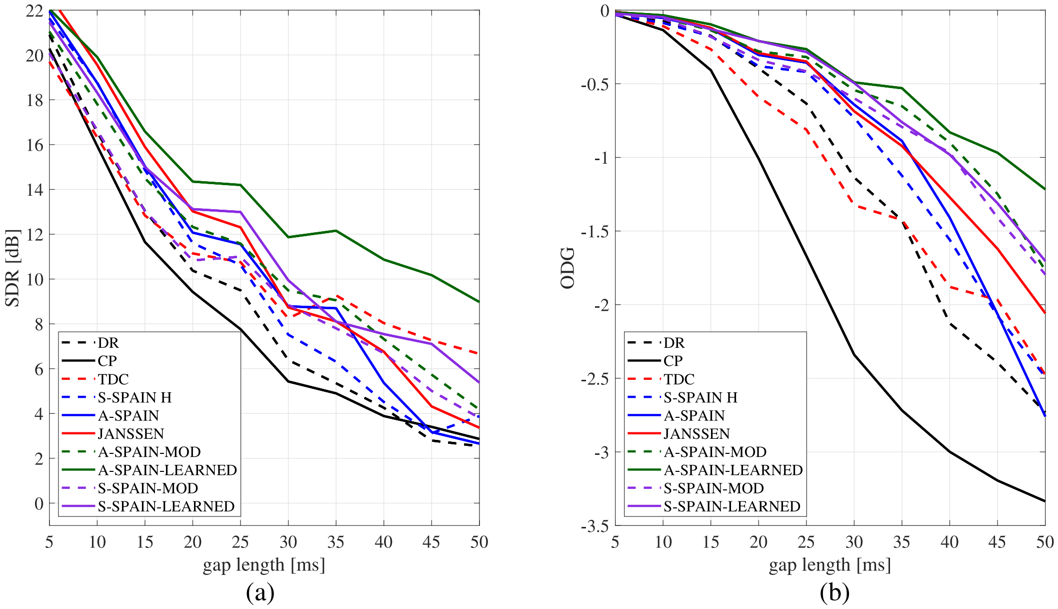


Fig. 7. Overall performance comparison of various audio inpainting algorithms. (a) depicts results in terms of SDR, (b) in terms of ODG values.

S-SPAIN-LEARNED outperform all other inpainting methods essentially over the whole gap length range (5 ms – 50 ms). Among those four, A-SPAIN-LEARNED, i.e., the analysis variant with learned dictionary, achieves best performance. We also observe that A-SPAIN-LEARNED is superior to any other algorithm in terms of SDR, thus, illustrating the large benefit of a sparsity-optimized dictionary. Similarly, S-SPAIN-LEARNED exhibits substantial improvements over S-SPAIN-MOD with respect to both SDR and ODR performance measures. Note, however, that these gains come at the expense of increased computational complexity due to the additional learning step.

Comparing analysis and synthesis model, we observe the common phenomenon that the analysis variants outperform the corresponding synthesis variants in terms of inpainting performance (with and without dictionary learning). It appears that the advantage of a sparsely synthesized original signal according to the synthesis model is outweighed by the fact that also more incorrect signals from the feasible set Γ_{x} can be sparsely synthesized.

VII. Software and Reproducible Research

The MATLAB codes needed for the experiments, all the data and supplemental figures are available at http://oeaw.ac.at/isf/dictlearnaudioinpaint.

VIII. Conclusion

We introduced a dictionary learning framework for audio inpainting. Our proposed method learns the dictionary from reliable parts around the gap such that a signal representation with increased sparsity is obtained. To that end, we developed a basis optimization technique to deform a given Gabor frame such that the sparsity of the analysis coefficients of the resulting frame is maximized. Since the optimization procedure is tailored to generate unitary matrices, "nice" properties of the Gabor frame (e.g., tightness or painless property) are preserved and are also present in the learned frame. This is important, since we used the analysis operator of the deformed frame as the sparsifying transform in a modified SPAIN algorithm, which benefits from such properties. This modified SPAIN algorithm 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.

The experimental results demonstrated that the devised dictionary learning approach yields large performance gains in combination with the modified analysis SPAIN variant and significantly outperforms any other inpainting method compared with. The required additional computational complexity for learning is, however, not negligible, so that implementations with reduced complexity are a topic for future investigations. In particular, dictionary learning implementations that avoid the replacement of the the ℓ_∞-component of the $\ell_{0,\infty}$-norm by a summation would be desirable, see also Remarks 2 and 3. While the `CVX` toolbox [61] allowed us to demonstrate the feasibility of our approach, other toolboxes (like, e.g., the `UNLocBoX` [64]) and/or direct `C/C++` coding are probably more adequate choices in this regard. Moreover, a suitable adaptation of the ADMM algorithm for solving the dictionary learning step by $\ell_{0,\infty}$-minimization (see Remark 2) seems to be a promising approach for follow-up research.

We also note that the presented dictionary learning framework seems to be applicable to other sparsity-based inpainting techniques apart from SPAIN as well. We expect that careful combinations with, e.g., (re-)weighted ℓ_1-minimization methods will also increase their performance. Finally, we would like to emphasize that our approach is general in the sense that one could replace the Gabor frame by any other frame (e.g., non-stationary Gabor frames [44]) satisfying the painless condition as long as a certain rectangular time-frequency structure is available. This leaves considerable room for further improvements.

Appendix

A. Simplification of y-Update (7b)

The fact that the y-update (7b) 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 is an immediate consequence of the following theorem.

Theorem 1: Let $\mathrm{A}:\mathbb{C}^N\to\mathbb{C}^P$ and $\mathrm{S}:\mathbb{C}^N\to\mathbb{C}^N$ denote analysis and frame operator of a frame satisfying the painless condition (2), respectively, and let $\Gamma_\mathrm{x}=\{\mathrm{y}\in\mathbb{R}^N:\mathrm{M_R y}=\mathrm{M_R x}\}$. Then, for any $\mathrm{z}\in\mathbb{C}^P$,

$$\arg\min_{\mathrm{y}\in\Gamma_\mathrm{x}}\|\mathrm{Ay}-\mathrm{z}\|_2 = \arg\min_{\mathrm{y}\in\Gamma_\mathrm{x}}\left\|\mathrm{y}-\mathrm{D}^{\,\mathrm{cd}}\mathrm{z}\right\|_2,$$

where $\mathrm{D}^{\,\mathrm{cd}}=(\mathrm{AS}^{-1})^\mathsf{H}$ denotes the synthesis operator of the canonical dual frame [35].

Proof: According to (2), $\mathrm{A^H A}=\mathrm{S}$ with diagonal S. The diagonal elements of S are strictly positive. Clearly, $(\mathrm{AS}^{-1/2})^\mathsf{H}(\mathrm{AS}^{-1/2})=\mathrm{S}^{-1/2}\mathrm{A^H A S}^{-1/2}=\mathrm{I}_N$, where $\mathrm{S}^{-1/2}$ is the diagonal $N\times N$ matrix, whose diagonal elements are the reciprocals of the (positive) square roots of the diagonal elements of S. Therefore

$$\mathrm{AS}^{-1/2}=\mathrm{U}\begin{bmatrix}\mathrm{I}_N\\0_{(P-N)\times N}\end{bmatrix}\mathrm{V}^\mathsf{H},$$

with $\mathrm{U}\in\mathbb{C}^{P\times P}$ and $\mathrm{V}\in\mathbb{C}^{N\times N}$ unitary, represents the singular value decomposition (SVD) of $\mathrm{AS}^{-1/2}$ [59]. Furthermore, let $\widetilde{\mathrm{U}}\in\mathbb{C}^{P\times P}$ denote the unitary matrix that is obtained from U by multiplying the first N columns with V^H and leaving the remaining columns unchanged, i.e.,

$$\widetilde{\mathrm{U}}=\mathrm{U}\begin{bmatrix}\mathrm{V}^\mathsf{H} & 0_{N\times(P-N)}\\0_{(P-N)\times N} & \mathrm{I}_{(P-N)}\end{bmatrix}.$$

Then,

$$\mathrm{AS}^{-1/2}=\widetilde{\mathrm{U}}\begin{bmatrix}\mathrm{I}_N\\0_{(P-N)\times N}\end{bmatrix},$$

so that

$$\mathrm{A}=\widetilde{\mathrm{U}}\begin{bmatrix}\mathrm{S}^{1/2}\\0_{(P-N)\times N}\end{bmatrix}. \tag{22}$$

Therefore,

$$\begin{aligned}\hat{\mathrm{y}}&=\arg\min_{\mathrm{y}\in\Gamma_\mathrm{x}}\|\mathrm{Ay}-\mathrm{z}\|_2^2\\&=\arg\min_{\mathrm{y}\in\Gamma_\mathrm{x}}\left\|\widetilde{\mathrm{U}}\begin{bmatrix}\mathrm{S}^{1/2}\\0_{(P-N)\times N}\end{bmatrix}\mathrm{y}-\mathrm{z}\right\|_2^2\\&=\arg\min_{\mathrm{y}\in\Gamma_\mathrm{x}}\left\|\begin{bmatrix}\mathrm{S}^{1/2}\\0_{(P-N)\times N}\end{bmatrix}\mathrm{y}-\widetilde{\mathrm{U}}^\mathsf{H}\mathrm{z}\right\|_2^2\\&=\arg\min_{\mathrm{y}\in\Gamma_\mathrm{x}}\left\|\mathrm{S}^{1/2}\mathrm{y}-\widetilde{\mathrm{U}}^\mathsf{H}_{\text{top}}\mathrm{z}\right\|_2^2,\end{aligned}$$

where $\widetilde{\mathrm{U}}^\mathsf{H}_{\text{top}}$ is the submatrix of $\widetilde{\mathrm{U}}^\mathsf{H}$ consisting of its first N rows. Here, the last step follows from the observation that the arg min operation does not depend on the bottom block (corresponding to the remaining $P-N$ rows).

Now, note that any $\mathrm{y}\in\Gamma_\mathrm{x}$ can be uniquely decomposed into a part $\mathrm{y}^{\text{gap}}\in\mathbb{R}^N$ supported on the gap and the reliable part $\mathrm{M_R x}$ supported on the complement of the gap according to $\mathrm{y}=\mathrm{y}^{\text{gap}}+\mathrm{M_R x}$. Since $\mathrm{S}^{1/2}$ is diagonal and any $\mathrm{y}\in\Gamma_\mathrm{x}$ is real-valued, this implies that

$$\hat{\mathrm{y}}_n=\begin{cases}(\mathrm{M_R x})_n, & n \text{ outside gap}\\\left(\mathrm{Re}\left(\mathrm{S}^{-1/2}\widetilde{\mathrm{U}}^\mathsf{H}_{\text{top}}\mathrm{z}\right)\right)_n, & \text{otherwise}\end{cases}. \tag{23}$$

Finally, we have

$$\begin{aligned}\mathrm{S}^{-1/2}\widetilde{\mathrm{U}}^\mathsf{H}_{\text{top}}&=\left[\mathrm{S}^{-1/2}\;\;0_{N\times(P-N)}\right]\widetilde{\mathrm{U}}^\mathsf{H}\\&=\left(\widetilde{\mathrm{U}}\begin{bmatrix}\mathrm{S}^{-1/2}\\0_{(P-N)\times N}\end{bmatrix}\right)^\mathsf{H}=(\mathrm{AS}^{-1})^\mathsf{H},\end{aligned}$$

see (22) for the last step, which – together with (23) – proves the claim. ■

B. Approximation of B*-Update (9a)*

We aim at justifying that for any $\mathrm{v} \in \mathbb{R}^N$,

$$\arg\min_{\mathrm{B}:\|\mathrm{B}\|_{0,\infty}\le k} \left\|\mathrm{D}\,\eta^\dagger\left(\mathrm{B}\right) - \mathrm{v}\right\|_2 \quad \approx \quad \mathcal{H}_k^{\mathrm{TF}}\left(\eta\left(\mathrm{A}^{\mathrm{cd}}\mathrm{v}\right)\right),$$

where $\mathrm{D} : \mathbb{C}^P \to \mathbb{C}^N$ denotes the synthesis operator of a frame and $\mathrm{A}^{\mathrm{cd}} = \mathrm{D}^{\mathsf{H}}\mathrm{S}^{-1}$ is the analysis operator of its canonical dual frame [35]. Since $\mathrm{D}\mathrm{A}^{\mathrm{cd}} = \mathrm{I}_N$, we have

$$\begin{aligned} \left\|\mathrm{D}\,\eta^\dagger\left(\mathrm{B}\right) - \mathrm{v}\right\|_2^2 &= \left\|\mathrm{D}\,\eta^\dagger\left(\mathrm{B}\right) - \mathrm{D}\mathrm{A}^{\mathrm{cd}}\mathrm{v}\right\|_2^2 \\ &= \left\|\mathrm{D}\left(\eta^\dagger\left(\mathrm{B}\right) - \mathrm{A}^{\mathrm{cd}}\mathrm{v}\right)\right\|_2^2 \\ &\le K_{\text{upper}}\left\|\eta^\dagger\left(\mathrm{B}\right) - \mathrm{A}^{\mathrm{cd}}\mathrm{v}\right\|_2^2, \qquad (24) \end{aligned}$$

where K_{upper} is an upper frame bound [35] of the given frame. The upper bound (24) is minimized for $\hat{\mathrm{B}} = \mathcal{H}_k^{\mathrm{TF}}(\eta(\mathrm{A}^{\mathrm{cd}}\mathrm{v}))$, so that we expect $\|\mathrm{D}\,\eta^\dagger(\hat{\mathrm{B}}) - \mathrm{v}\|_2^2$ to be sufficiently close to the true minimum $\min_{\mathrm{B}:\|\mathrm{B}\|_{0,\infty}\le k}\|\mathrm{D}\,\eta^\dagger(\mathrm{B}) - \mathrm{v}\|_2$.

Acknowledgment

The authors would like to thank N. Holighaus for helpful discussions. They are also grateful to the anonymous reviewers for numerous constructive comments that have resulted in a major improvement of this paper.

References

[1] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley, "Audio inpainting," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

[2] S. Godsill, P. Rayner, and O. Cappé, "Digital audio restoration," in *Proc. Appl. Digit. Signal Process. Audio Acoust.* Springer, 2002, pp. 133–194.

[3] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Netw.*, vol. 12, no. 5, pp. 40–48, Sep./Oct. 1998.

[4] Y. Bahat, Y. Y. Schechner, and M. Elad, "Self-content-based audio inpainting," *Signal Process.*, vol. 111, pp. 61–72, 2015.

[5] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 2, pp. 317–330, Apr. 1986.

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

[7] W. Etter, "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Trans. Signal Process.*, vol. 44, no. 5, pp. 1124–1135, May 1996.

[8] P. A. Esquef, V. Välimäki, K. Roth, and I. Kauppinen, "Interpolation of long gaps in audio signals using the warped Burg's method," in *Proc. 6th Int. Conf. Digit. Audio Effects (DAFx)*, 2003, pp. 8–11.

[9] I. Kauppinen and K. Roth, "Audio signal extrapolation-theory and applications," in *Proc. 5th Int. Conf. Digit. Audio Effects (DAFx)*, 2002, pp. 105–110.

[10] I. Kauppinen and J. Kauppinen, "Reconstruction method for missing or damaged long portions in audio signal," *J. Audio Eng. Soc.*, vol. 50, no. 7/8, pp. 594–602, 2002.

[11] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, vol. 49, no. 3, pp. 906–931, Jun. 1989.

[12] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[13] E. J. Candés, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[14] P. Balazs, M. Doerfler, M. Kowalski, and B. Torresani, "Adapted and adaptive linear time-frequency representations: A synthesis point of view," *IEEE Signal Process. Mag.*, vol. 30, no. 6, pp. 20–31, Nov. 2013.

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

[16] F. Lieb and H.-G. Stark, "Audio inpainting: Evaluation of time-frequency representations and structured sparsity approaches," *Signal Process.*, vol. 153, pp. 291–299, 2018.

[17] O. Mokrý, P. Záviška, P. Rajmic, and V. Veselý, "Introducing SPAIN (SParse audio INpainter)," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[18] O. Mokrý and P. Rajmic, "Audio inpainting: Revisited and reweighted," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2906–2918, 2020.

[19] M. Lagrange, S. Marchand, and J.-B. Rault, "Long interpolation of audio signals using linear prediction in sinusoidal modeling," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 891–905, 2005.

[20] J. Lindblom and P. Hedelin, "Packet loss concealment based on sinusoidal modeling," in *Proc. Speech Coding, 2002, IEEE Workshop*, 2002, pp. 65–67.

[21] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs, "Inpainting of long audio segments with similarity graphs," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1083–1094, Jun. 2018.

[22] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A context encoder for audio inpainting," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2362–2372, Dec. 2019.

[23] A. Marafioti, N. Holighaus, P. Majdak, N. Perraudin *et al.* "Audio inpainting of music by means of neural networks," in *Proc. Audio Eng. Soc. Conv.*, Mar. 2019.

[24] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "A constrained matching pursuit approach to audio declipping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 329–332.

[25] B. Defraene, N. Mansour, S. De Hertogh, T. Van Waterschoot, M. Diehl, and M. Moonen, "Declipping of audio signals using perceptual compressed sensing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 12, pp. 2627–2637, Dec. 2013.

[26] K. Siedenburg, M. Kowalski, and M. Doerfler, "Audio declipping with social sparsity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1577–1581.

[27] Ç. Bilen, A. Ozerov, and P. Pérez, "Audio declipping via nonnegative matrix factorization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.

[28] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: A flexible non-convex approach," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Springer, 2015, pp. 243–250.

[29] F. R. Ávila, M. P. Tcheou, and L. W. Biscainho, "Audio soft declipping based on constrained weighted least squares," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1348–1352, Sep. 2017.

[30] L. Rencker, F. Bach, W. Wang, and M. D. Plumbley, "Consistent dictionary learning for signal declipping," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*. Springer, 2018, pp. 446–455.

[31] P. Záviška, P. Rajmic, Z. Průša, and V. Veselý, "Revisiting synthesis model in sparse audio declipper," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*. Springer, 2018, pp. 429–445.

[32] P. Záviška, P. Rajmic, O. Mokrý, and Z. Průša, "A proper version of synthesis-based sparse audio declipper," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, 2019, pp. 591–595.

[33] P. Rajmic, P. Záviška, V. Veselý, and O. Mokrý, "A new generalized projection and its application to acceleration of audio declipping," *Axioms*, vol. 8, no. 3, p. 105, 2019.

[34] G. Tauböck, S. Rajbamshi, and P. Balazs, "Sparse audio inpainting: A dictionary learning technique to improve its performance," in *Audio Eng. Soc. Conv. 149*, Oct. 2020. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20939

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

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

[37] V. Mach and R. Ozdobinski, "Optimizing dictionary learning parameters for solving audio inpainting problem," *Int. J. Adv. Telecommun., Electrotechnics, Signals Syst.*, vol. 2, no. 1, pp. 39–44, 2013.

[38] C. Guichaoua, "Dictionary learning for audio inpainting," *HAL Robot.*, 2012.

[39] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[40] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1999, vol. 5, pp. 2443–2446.

[41] K. Schnass and F. Teixeira, "Compressed dictionary learning," *J. Fourier Anal. Appl.*, vol. 26, no. 2, pp. 1–37, 2020.

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

[43] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, no. 5, pp. 1271–1283, 1986.

[44] P. Balazs, M. Doerfler, F. Jaillet, N. Holighaus, and G. Velasco, "Theory, implementation and applications of nonstationary gabor frames," *J. Comput. Appl. Math.*, vol. 236, no. 6, pp. 1481–1496, 2011.

[45] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.

[46] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc.*, vol. 58, pp. 267–288, 1994.

[48] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[49] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Problems*, vol. 23, no. 3, p. 947, 2007.

[50] F. Krahmer, G. E. Pfander, and P. Rashkov, "Uncertainty in time-frequency representations on finite abelian groups and applications," *Appl. Comput. Harmonic Anal.*, vol. 25, no. 2, pp. 209–225, 2008.

[51] L. D. Abreu and M. Speckbacher, "Donoho-Logan large sieve principles for modulation and polyanalytic Fock spaces," 2018, *arXiv:1808.02258*.

[52] S. Boyd *et al.* "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[53] P. L. Søndergaard, B. Torrésani, and P. Balazs, "The linear time frequency analysis toolbox," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 10, no. 04, 2012, Art. no. 1250032.

[54] Z. Pruša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The large time-frequency analysis toolbox 2.0," in Sound, Music, and Motion, ser. Lncs. Springer International Publishing, 2014, pp. 419–442.

[55] P. Záviška, O. Mokrý, and P. Rajmic, "S-SPADE done right: Detailed study of the sparse audio declipper algorithms," Tech. Rep., Brno Univ. Tech., Sep. 2018, *arXiv:1809.09847*.

[56] G. Tauböck and F. Hlawatsch, "Compressed sensing based estimation of doubly selective channels using a sparsity-optimized basis expansion," in *Proc. 16th Eur. Signal Process. Conf.*, Lausanne, Switzerland, Aug. 2008, pp. 1–5.

[57] G. Tauböck, F. Hlawatsch, D. Eiwen, and H. Rauhut, "Compressive estimation of doubly selective channels in multicarrier systems: Leakage effects and sparsity-enhancing processing," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 255–271, Apr. 2010.

[58] R. Bhatia, *Matrix Analysis*. Springer, 1997.

[59] G. H. Golub and C. F. Van Loan, *Matrix Computations*, *3rd ed*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[60] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge (U.K.): Cambridge University Press, Dec. 2004.

[61] M. Grant and S. Boyd, CVX: Matlab Software for Disciplined Convex Programming (Web Page and Software), Stanford University, CA, USA. [Online]. Available: http://cvxr.com/cvx/

[62] R. Huber and B. Kollmeier, "PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[63] "EBU SQAM CD: Sound Quality Assessment Material Recordings for Subjective Tests." [Online]. Available: https://tech.ebu.ch/publications/sqamcd

[64] N. Perraudin, D. Shuman, G. Puy, and P. Vandergheynst, "UNLocBoX A Matlab convex optimization toolbox using proximal splitting methods," Feb. 2014, *arXiv:1402.0779*.