Submitted to the 23rd International Conference on Digital Audio Effects (DAFx-20), Vienna, Austria, September 8–12, 2020

FLEXIBLE FRAMEWORK FOR AUDIO RESTORATION

Ondřej Mokrý, Pavel Rajmic , Pavel Záviška*

Signal Processing Laboratory
Brno University of Technology, Brno, Czech Republic
rajmic@feec.vutbr.cz

ABSTRACT

The paper presents a unified, flexible framework for the tasks of audio inpainting, declipping, and dequantization. The concept is further extended to cover analogous degradation models in a transformed domain, e.g. quantization of the time-frequency coefficients. The problem of restoring an audio signal from degraded observations in two different domains is formulated as an inverse problem, and several algorithmic solutions are developed. The viability of the presented concept is demonstrated on an example where audio restoration from partial and quantized observations of both the time-domain signal and its time-frequency coefficients is carried on.

1. INTRODUCTION

Audio inpainting, audio declipping and audio dequantization are restoration tasks that are usually studied separately in the literature. In audio inpainting, some of the time-domain signal samples are completely missing and they need to be recovered, while in the cases of declipping and dequantization, the samples are not lost fully and the samples to be recovered are known to lie in prescribed numerical ranges, depending on the model of the degradation.

A unification of different kinds of audio restoration tasks has been partially discussed in [1], where the authors covered dequantization and declipping (possibly at the same time), and in [2, 3], whose formulation allowed denoising and declipping (but not simultaneously). A flexible algorithmic framework is presented also in [4], based on the non-negative matrix factorization (shown to be suitable for simultaneous audio declipping and click concealment). The present article shows how the three tasks can be covered by a unified restoration framework, all of them possibly taking effect at the same time. The greatest contrast to the earlier attempts is however that this paper extends the range of degradation models by additionally moving to a transformed domain, i.e. missing, clipped and quantized observations are further allowed after (linearly) transforming the signal, e.g. by the short-time Fourier transform.

In Section 2, we introduce the three respective audio degradation models in more detail, emphasize their common factors, and build the set of feasible time-domain signals, which contains the potential solutions to the recovery task. We then extend the degradation to the transformed domain and present the synthesis and analysis variants of the resulting feasible set.

* Corresponding author.

Finding a solution of any of the described recovery tasks is generally ill-posed. A regularizer is needed to pick favorable candidates from the feasibility set. The sparsity of time-frequency-transformed audio signals has been shown to be a suitable regularizer for audio recovery problems [5, 6, 7, 8]. Thus, Section 3 presents the general optimization problem with a special emphasis on using ℓ_1 relaxation of true sparsity. The section also presents a single, unified algorithm to find the numerical solution in the case of a convex regularizer.

In Section 4, we present a proof-of-concept example of an audio codec (i.e., coder and decoder). In the coder part, the original, input audio signal is due to subsampling *and* quantization in both time *and* time-frequency domains. The decoder attempts at recovering the signal from this partial information, based on the assumption of sparsity of the (now unknown) original. Today's audio codecs are built on the single-domain information, for instance the classical MPEG model codes the TF coefficients only, based on the global masking threshold estimate [9]. Recovery from quantized transformed observations is also studied in [10] in the context of compressed sensing. The interesting recent approach from [11], which is inspired in the image processing field, subsamples and quantizes purely time-domain audio samples to achieve compression. We show experimentally that in contrast to that approach, splitting the available bit budget between the two domains can be beneficial in some cases.

2. BUILDING THE FRAMEWORK

2.1. Time-frequency representations

In audio processing, time-frequency (TF) operators are usually used [12] to provide a suitable representation of a signal. A signal $\mathbf{x} \in \mathbb{R}^P$ is represented as a superposition of time-localized oscillations, where the localization is due to the so-called window function that moves along the signal. Among such TF operators, the so-called *tight frames* are usually preferred, since they provide effective handling of both theoretical derivations and practical computations [12, 13, 14]. The Short-time Fourier (STFT) or the Modified Discrete Cosine (MDCT) transforms [5, 15] are classical examples of such operators.

Throughout the paper, we use the following convention: To obtain an expansion of a signal $\mathbf{x} \in \mathbb{R}^P$ to a series of TF coefficients, the *analysis* operator $A\colon \mathbb{R}^P \to \mathbb{C}^Q$ is applied, where we assume $Q \geq R$. The adjoint, *synthesis* operator $A^*\colon \mathbb{C}^Q \to \mathbb{R}^P$, reproduces the time-domain signal from the coefficients.

A tight frame with frame bound α can be characterized by the property $A^*A = \alpha\mathrm{Id}$, where Id is the identity operator, here on the space $\mathbb{C}^Q$. Furthermore, denote $\mathcal{R}(A) \subset \mathbb{C}^Q$ the range space of the analysis operator, i.e. the set of spectra consistent with the time-domain signals. In case of a tight frame with frame bound α, the orthogonal projection onto $\mathcal{R}(A)$ is simply expressed as

$\mathrm{proj}_{\mathcal{R}(A)} = \alpha^{-1}AA^*$ [16]. When the constant $\alpha = 1$, the frame is said to be Parseval tight.

2.2. Inpainting

Audio inpainting is a general term for recovering missing or highly degraded samples of the audio signal [5]. Suppose $\mathbf{y} \in \mathbb{R}^P$ is the original, non-degraded signal, and $M\mathbf{y}$ is the partial observation of $\mathbf{y}$. The operator $M\colon \mathbb{R}^P \to \mathbb{R}^P$ keeps the reliable samples, while putting zeros at the positions of missing or unreliable samples; these positions are assumed to be known. Thus, M can be identified with a diagonal matrix $\mathbf{M}$ of size $P \times P$, for which $m_{pp} = 1$ for a reliable sample y_p, zero otherwise. The solution of the inpainting problem is called *consistent* if it lies in a naturally defined set

$$\Gamma_{\mathrm{T}}^{\mathrm{inp}} = \left\{ \mathbf{x} \in \mathbb{R}^P \mid M\mathbf{x} = M\mathbf{y} \right\}. \tag{1}$$

Clearly, defining the consistency set alone is not sufficient to solve the inpainting problem, since the inverse problem is ill-posed. Thus, the path to the solution must start from a careful consideration of additional assumptions about the signal. To name but a few, the solution may be modeled as an autoregressive process [17, 18], as a sum of sinusoidal components [19], or it is assumed to be sparse with respect to a suitable TF transform [5, 7, 20].

For the purpose of further generalization, the time-domain set $\Gamma_{\mathrm{T}}^{\mathrm{inp}}$ may be equivalently defined entrywise as a box-type set

$$\mathbf{x} \in \Gamma_{\mathrm{T}}^{\mathrm{inp}} \;\Leftrightarrow\; \begin{cases} x_p \in [y_p, y_p] \text{ for reliable indexes } p, \\ x_p \in (-\infty, +\infty) \text{ otherwise,} \end{cases} \tag{2}$$

for $p = 1, 2, \ldots, P$.

2.3. Declipping

Audio declipping aims at recovering a signal damaged by clipping. This negative effect is one of the common audio degradation types and it can be described as a non-linear distortion causing a limitation of a signal, such that all values of the signal in waveform exceeding the allowed dynamic range defined by thresholds $[-\theta, \theta]$ are strictly limited to these thresholds. Because of the strict limitation of signal samples, the effect is also referred to as *hard clipping*. Not only is the information contained in the peaks lost, but clipping also introduces a great number of higher harmonics, which leads to a significant reduction of the perceived audio quality [21] and also the accuracy of automatic voice recognition [22].

Audio declipping is similar to audio inpainting, with the difference that additional information about the lower and upper bounds for values of the declipped samples is available. Actually, simple inpainting methods are able to effectively perform declipping, such as the Janssen method used in [5]. In general, however, inpainting approaches to declipping clearly break the *consistency* requirement of the solution.

Similarly to the inpainting case, the set of consistent solutions, $\Gamma_{\mathrm{T}}^{\mathrm{dec}}$, is defined entrywise, taking advantage of the information that declipped samples need to exceed the clipping thresholds:

$$\mathbf{x} \in \Gamma_{\mathrm{T}}^{\mathrm{dec}} \;\Leftrightarrow\; \begin{cases} x_p \in [y_p, y_p] \text{ for reliable samples } y_p, \\ x_p \in (-\infty, -\theta] \text{ for observed samples } -\theta, \\ x_p \in [\theta, +\infty) \text{ for observed samples } \theta. \end{cases} \tag{3}$$

2.4. Dequantization

The term dequantization refers to an inverse problem where a signal should be recovered based on the knowledge of its quantized version. In this subsection, the quantization acts in the time domain, i.e. directly on the audio samples; the original sample is substituted with the value of the nearest quantization level. The unique quantization level is identified using a pair of the nearest so-called decision levels [23].

More specifically, assume a series of quantization levels

$$\cdots < q_{-1} < q_0 < q_1 < q_2 < \ldots. \tag{4}$$

where this sequence can be infinite (in theory, but always finite in practice). Fix p for the moment. For the input sample y_p there exist a unique n such that it holds $q_n \leq y_p < q_{n+1}$. Based on the decision level d_n for which $q_n < d_n < q_{n+1}$, quantization maps y_p either to q_n (when $y_p < d_n$) or to q_{n+1} (when $y_p \geq d_n$). In turn, if a quantized value y_p^{quant} is observed, there exist a single interval $[d_n, d_{n+1})$ to which y_p belonged.

Therefore, for the purpose of formulating the general problem, the set of solutions consistent with the quantization model is defined again as the box-type set $\Gamma_{\mathrm{T}}^{\mathrm{deq}}$,

$$\mathbf{x} \in \Gamma_{\mathrm{T}}^{\mathrm{deq}} \;\Leftrightarrow\; x_p \in [d_n, d_{n+1}), \tag{5}$$

where d_n, d_{n+1} changes depending on p, which is intentionally not reflected by the notation. Note also that in the finite case, border cases can be treated by using $\pm\infty$ in place of lower or upper bound in (5). In such a case, the closed interval should be apparently replaced by an open one.

2.5. General formulation

When working with digitized signals, declipping can actually be seen as a special kind of quantization, where the set of quantization levels defined by Eq. (4) cover precisely the range of possible values in the range $[-\theta, \theta]$.

Even more, looking at definitions (2), (3) and (5), one may observe that it is straightforward to define a feasible set for *simultaneous* audio inpainting, declipping and dequantization. Such a set is defined entrywise as a multidimensional interval Γ_{T} such that

$$\mathbf{x} \in \Gamma_{\mathrm{T}} \;\Leftrightarrow\; x_p \in [l_{\mathrm{T}p}, u_{\mathrm{T}p}], \tag{6}$$

where the entries of the vector lower bound $\mathbf{l}_{\mathrm{T}}$ and the upper bound $\mathbf{u}_{\mathrm{T}}$ depend on the type of degradation that occurs at index p, $p = 1, \ldots, P$. Recall that the bounds may formally contain plus or minus infinity.

It is straightforward to show that the set Γ_{T} is convex. Furthermore, solving an inverse problem with such a set of feasible solutions is viable since projection onto this type of set is available explicitly and entrywise by

$$\left(\mathrm{proj}_{\Gamma_{\mathrm{T}}}(\mathbf{x})\right)_p = \min\left\{u_{\mathrm{T}p}, \max\left\{x_p, l_{\mathrm{T}p}\right\}\right\}. \tag{7}$$

2.6. Using the information of two different domains

So far, only time-domain degradation has been considered, leading to the set Γ_{T}. Nevertheless, degradation as presented above can also happen in a transformed domain. The aim of this section is to generalize the above concept to both time and time-frequency domains.

Similarly to (6), we define a *consistent* feasible set within the TF domain. Such a domain is generally a subset of $\mathbb{C}^Q$, and therefore any interval shall be understood in such a way that the real and imaginary parts are considered independently. As an example, for $l, u \in \mathbb{C}$, we denote

$$z \in [l, u] \;\Leftrightarrow\; \Re(z) \in [\Re(l), \Re(u)] \wedge \Im(z) \in [\Im(l), \Im(u)]. \quad (8)$$

With such a notation, we define the membership in Γ_{TF} as

$$\mathbf{z} \in \Gamma_{\mathrm{TF}} \;\Leftrightarrow\; z_q \in [l_{\mathrm{TF}q}, u_{\mathrm{TF}q}], \quad (9)$$

for all $q = 1, 2, \ldots, Q$. The vectors $\mathbf{l}_{\mathrm{TF}}, \mathbf{u}_{\mathrm{TF}} \in \mathbb{C}^Q$ determine for each coefficient whether its clipped or quantized version is observed, or the coefficient is missing.

2.7. Involving a prior

Combining constraints in the two domains reduces the size of the feasible set $\Gamma_{\mathrm{T}} \cap \Gamma_{\mathrm{TF}}$, which is valuable for finding the restored signal. In general, however, this is not enough, and additional prior information is necessary. For the purpose of our general framework, assume knowledge about the TF coefficients invoked by minimizing a functional $\mathcal{S} \circ W$. As a particular example, consider the (relaxed) sparse prior, $(\mathcal{S} \circ W)(\mathbf{z}) = \mathcal{S}(W\mathbf{z}) = \|W\mathbf{z}\|_1$, where W is a diagonal operator assigning weights to the respective coefficients. The ℓ_1 norm sums the magnitudes of elements of its argument [24].

Combining the prior and the feasible sets Γ_{T} and Γ_{TF} provides us with the following general formulation:

$$\arg\min_{\mathbf{u}} \mathcal{S}(WK\mathbf{u}) \quad \text{subject to} \quad L\mathbf{u} \in \Gamma_{\mathrm{T}},\ K\mathbf{u} \in \Gamma_{\mathrm{TF}}. \quad (10)$$

The linear operators K and L play the role of either synthesis or analysis operator of a suitable TF transform. In a typical situation, one of them will be identity. Such a notation may look redundant, but the reason for this shape of the formulation (10) is that it covers both the synthesis formulation

$$\arg\min_{\mathbf{z}} \mathcal{S}(W\mathbf{z}) \quad \text{subject to} \quad A^*\mathbf{z} \in \Gamma_{\mathrm{T}},\ \mathbf{z} \in \Gamma_{\mathrm{TF}}, \quad (11)$$

when K is the identity, $K = \mathrm{Id}$, and the analysis formulation in the case $L = \mathrm{Id}$:

$$\arg\min_{\mathbf{x}} \mathcal{S}(WA\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \Gamma_{\mathrm{T}},\ A\mathbf{x} \in \Gamma_{\mathrm{TF}}. \quad (12)$$

Note that once a non-unitary transform A is used, the formulations (11) and (12) are not equivalent.

3. SOLVING THE GENERAL TASK

The important observation about the sets Γ_{T} and Γ_{TF} defined by (6) and (9), respectively, is that both are box-type, thus convex sets. Furthermore, both the sets $\Gamma_L = \{\mathbf{u} \mid L\mathbf{u} \in \Gamma_{\mathrm{T}}\}$ and $\Gamma_K = \{\mathbf{u} \mid K\mathbf{u} \in \Gamma_{\mathrm{TF}}\}$ are convex as well. The reason is that the preimage of a convex set under a linear operator is a convex set, which is straightforward to show. Finally, the intersection of two convex sets is once again a convex set, therefore the set of feasible solutions in the constrained problem (10) is convex for arbitrary linear operators L and K.

However, such an intersection is a rather complicated set. One of the sets Γ_L and Γ_K is no more a box-type set, hence the intersection $\Gamma_L \cap \Gamma_K$ is generally a polyhedron either in the time domain (for the analysis model, $K = A$, $L = \mathrm{Id}$) or in the TF domain (for the synthesis model, $K = \mathrm{Id}$, $L = A^*$). This difficulty is treated right in the following section.

Based on the above observations, Sections 3.1 and 3.2 mainly focus on the convex setting. Sections 3.3, 3.4 and 3.5 will suggest alternative approaches; however, those will not be further developed in the present paper.

3.1. Consistent convex approach, arbitrary linear operators

We first focus on the case when the function $\mathcal{S}$ is convex, thus the whole general problem is convex. The idea is to use the proximal splitting [25] to solve the general formulation (10) numerically, which allows us to focus separately on $\mathcal{S}$ and the two constraints $\mathbf{u} \in \Gamma_L$ and $\mathbf{u} \in \Gamma_K$. We further utilize the possibility to efficiently compute the projection of a vector onto a box-type set. This makes sense especially when the proximal operator of $\mathcal{S}$ is assumed to have an explicit form, which will be the case below.

Note that the proximal operator of a convex lower semi-continuous function $h\colon \mathbb{V} \to \mathbb{R}$, denoted $\mathrm{prox}_h\colon \mathbb{V} \to \mathbb{V}$, is defined as $\mathrm{prox}_h(\mathbf{u}) = \arg\min_{\mathbf{v}} \left\{h(\mathbf{v}) + \frac{1}{2}\|\mathbf{v} - \mathbf{u}\|^2\right\}$ at any point $\mathbf{u} \in \mathbb{V}$. Here, $\mathbb{V}$ stands for the Hilbert space $\mathbb{C}^P$ or $\mathbb{C}^Q$. We assume that when prox_h is available, also $\mathrm{prox}_{\gamma h}$ is available for an arbitrary constant $\gamma > 0$.

To design a particular proximal algorithm, the problem (10) is first rewritten into the unconstrained form using the so-called *indicator function* ι_Γ of the set Γ. For $\mathbf{u} \in \Gamma$, the function returns 0, and ∞ otherwise. The general unconstrained formulation (10) thus attains the form

$$\arg\min_{\mathbf{u}} \left\{\mathcal{S}(WK\mathbf{u}) + \iota_{\Gamma_{\mathrm{T}}}(L\mathbf{u}) + \iota_{\Gamma_{\mathrm{TF}}}(K\mathbf{u})\right\}. \quad (13)$$

Such a form is suitable for the use of the generic proximal algorithm proposed independently by Condat [26] and Vũ [27] (further referred to as the CV algorithm). It is tailored to solve the problems of the form

$$\arg\min_{\mathbf{u}} \left\{f(\mathbf{u}) + g(\mathbf{u}) + \sum_{m=1}^{M} h_m(L_m\mathbf{u})\right\}, \quad (14)$$

where $f, g, h_1, \ldots, h_m$ are convex lower semi-continuous functions and f is differentiable. We will utilize the second of the two proposed variants from [26], the general form of which is reproduced in Alg. 1.

Assuming a finite-dimensional problem together with $f = 0$, the sequence $(\mathbf{u}^{(i)})_{i\in\mathbb{N}}$ produced by the algorithm is guaranteed to converge to the solution of problem (14) if

$$\tau\sigma\|\sum_{m=1}^{M} L_m^* L_m\| \le 1, \quad 0 < \rho < 2. \quad (15)$$

To develop the case-specific form of Alg. 1, the functions from formulation (13) are assigned as follows:

$$h_1 = \mathcal{S}, \qquad h_2 = \iota_{\Gamma_{\mathrm{T}}}, \qquad h_3 = \iota_{\Gamma_{\mathrm{TF}}}, \quad (16)$$

$$L_1 = WK, \qquad L_2 = L, \qquad L_3 = K, \quad (17)$$

and the functions f, g are both zero. Finally, we apply the following general properties:

- Since $g = 0$, it holds $\mathrm{prox}_{\tau g} = \mathrm{Id}$.

Algorithm 1: The CV algorithm for solving (14)

Input: The linear operators $L_m,\ m = 1, \dots, M$, the proximal operators $\mathrm{prox}_{h_m},\ m = 1, \dots, M$, prox_g and the gradient ∇f.

Choose the parameters $\tau, \sigma, \rho > 0$.

Choose the initial estimates $\mathbf{u}^{(0)}, \mathbf{v}_1^{(0)}, \dots, \mathbf{v}_M^{(0)}$.

for $i = 0, 1, \dots$ **do**

  for $m = 1, \dots, M$ **do**

    $\tilde{\mathbf{v}}_m^{(i+1)} = \mathrm{prox}_{\sigma h_m^*}\left(\mathbf{v}_m^{(i)} + \sigma L_m \mathbf{u}^{(i)}\right)$

    $\mathbf{v}_m^{(i+1)} = \rho \tilde{\mathbf{v}}_m^{(i+1)} + (1-\rho)\mathbf{v}_m^{(i)}$

  end

  $\tilde{\mathbf{u}}^{(i+1)} = \mathrm{prox}_{\tau g}\left(\mathbf{u}^{(i)} - \tau \nabla f\left(\mathbf{u}^{(i)}\right) - \tau \sum L_m^*\left(2\tilde{\mathbf{v}}_m^{(i+1)} - \mathbf{v}_m^{(i)}\right)\right)$

  $\mathbf{u}^{(i+1)} = \rho \tilde{\mathbf{u}}^{(i+1)} + (1-\rho)\mathbf{u}^{(i)}$

end

Output: $\mathbf{u}^{(i+1)}$

- To evaluate $\mathrm{prox}_{\sigma h^*}$, where h^* is the Fenchel–Rockafellar conjugate of h, one may benefit from the Moreau identity $\mathrm{prox}_{\sigma h^*}(\mathbf{u}) = \mathbf{u} - \sigma \mathrm{prox}_{h/\sigma}(\mathbf{u}/\sigma)$ [28].
- The proximal operator of an indicator function ι_Γ of a closed convex set Γ is the projection onto the set, proj_Γ.

Plugging these properties into Alg. 1 produces the algorithm for the general problem (13), and thus for (10). The final algorithm is summarized in Alg. 2.

The strength of the algorithm is that both the projections can be performed explicitly and fast, entry by entry. For the time-domain projection $\mathrm{proj}_{\Gamma_\mathrm{T}}$, Eq. (7) is used. For the TF-domain projection $\mathrm{proj}_{\Gamma_\mathrm{TF}}$, the same equation can be adapted, since the projection can be done not only entrywise, but also separately for the real and imaginary parts.

Note that the functions in problem (13) were assigned to the functions h_1, h_2, h_3 such that Alg. 2 covers both the synthesis and the analysis approaches (11) and (12), respectively. Had the composition $\mathcal{S} \circ K$ been assigned to the function g instead, the operator $\mathrm{prox}_{\tau g}$ would be known only in the synthesis model.[1]

3.2. Consistent convex approach, tight frame case

As mentioned in [26], if possible, one should make use of the functions f and g in Eq. (14) when assigning the functions of a particular problem. This is not possible in the case of the general formulation (13), unless the linear operators represent analysis or synthesis of a tight frame. In such a case, we may assign

$$g = \iota_{\Gamma_\mathrm{T}} \circ L, \qquad h_1 = \mathcal{S}, \qquad h_2 = \iota_{\Gamma_\mathrm{TF}}, \tag{18}$$

$$L_1 = WK, \qquad L_2 = K. \tag{19}$$

This is justified by the observation that in the case of a tight frame, L is either the synthesis (in the synthesis model), or the identity on the time domain (in the analysis model). In both cases, it satisfies $LL^* = \alpha \mathrm{Id}$ for a constant α, allowing us to compute the proximal operator $\mathrm{prox}_{\iota_{\Gamma_\mathrm{T}} \circ L}$ using the explicit formula [25, 29]

$$\mathrm{prox}_{\iota_{\Gamma_\mathrm{T}} \circ L}(\mathbf{u}) = \mathbf{u} + \alpha^{-1} L^*\left(\mathrm{proj}_{\Gamma_\mathrm{T}}(L\mathbf{u}) - L\mathbf{u}\right). \tag{20}$$

[1] The potential evaluation of $\mathrm{prox}_{\tau g} = \mathrm{prox}_{\tau \mathcal{S} \circ A}$ in the analysis model is complicated, because the formula for a proximal operator of such a composition is known only when the operator A satisfies $AA^* = \alpha \mathrm{Id}$, which is not possible in the setting of redundant TF transforms [29].

Algorithm 2: The CV algorithm for solving the general formulation (10)

Input: The linear operators W, K, L, the proximal operator $\mathrm{prox}_{\mathcal{S}}$ and the projectors $\mathrm{proj}_{\Gamma_\mathrm{T}}, \mathrm{proj}_{\Gamma_\mathrm{TF}}$.

Choose the parameters $\tau, \sigma, \rho > 0$.

Choose the initial estimates $\mathbf{u}^{(0)}, \mathbf{v}_1^{(0)}, \mathbf{v}_2^{(0)}, \mathbf{v}_3^{(0)}$.

for $i = 0, 1, \dots$ **do**

  /* update corresponding to h_1 */

  $\tilde{\mathbf{v}}_1^{(i+1)} = \mathbf{v}_1^{(i)} + \sigma WK\mathbf{u}^{(i)} - \sigma \mathrm{prox}_{\mathcal{S}/\sigma}\left(\mathbf{v}_1^{(i)}/\sigma + WK\mathbf{u}^{(i)}\right)$

  $\mathbf{v}_1^{(i+1)} = \rho \tilde{\mathbf{v}}_1^{(i+1)} + (1-\rho)\mathbf{v}_1^{(i)}$

  /* update corresponding to h_2 */

  $\tilde{\mathbf{v}}_2^{(i+1)} = \mathbf{v}_2^{(i)} + \sigma L\mathbf{u}^{(i)} - \sigma \mathrm{proj}_{\Gamma_\mathrm{T}}\left(\mathbf{v}_2^{(i)}/\sigma + L\mathbf{u}^{(i)}\right)$

  $\mathbf{v}_2^{(i+1)} = \rho \tilde{\mathbf{v}}_2^{(i+1)} + (1-\rho)\mathbf{v}_2^{(i)}$

  /* update corresponding to h_3 */

  $\tilde{\mathbf{v}}_3^{(i+1)} = \mathbf{v}_3^{(i)} + \sigma K\mathbf{u}^{(i)} - \sigma \mathrm{proj}_{\Gamma_\mathrm{TF}}\left(\mathbf{v}_3^{(i)}/\sigma + K\mathbf{u}^{(i)}\right)$

  $\mathbf{v}_3^{(i+1)} = \rho \tilde{\mathbf{v}}_3^{(i+1)} + (1-\rho)\mathbf{v}_3^{(i)}$

  /* update of $\mathbf{u}$ */

  $\mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} - \rho\tau K^* W^*\left(2\tilde{\mathbf{v}}_1^{(i+1)} - \mathbf{v}_1^{(i)}\right) - \rho\tau L^*\left(2\tilde{\mathbf{v}}_2^{(i+1)} - \mathbf{v}_2^{(i)}\right) - \rho\tau K^*\left(2\tilde{\mathbf{v}}_3^{(i+1)} - \mathbf{v}_3^{(i)}\right)$

end

Output: $\mathbf{u}^{(i+1)}$

Put into words, the formula states that instead of computing the complicated projection on the left-hand side, one may use the simple projection onto Γ_T on the right-hand side, together with the application of the linear operator and its adjoint.

The resulting algorithm is summarized by Alg. 3, where, for simplicity, $\alpha = 1$ is assumed (i.e., the frame is Parseval tight). Compared to Alg. 2, this has three major benefits:

- for $\rho \leq 1$, every iterate $\mathbf{u}^{(i+1)}$ lies in Γ_T,
- in [26], it is suggested that involving the function g may result in faster convergence of the algorithm,
- since it uses only two functions h_1, h_2 and thus only two corresponding linear operators, it follows from Eq. (15) that a wider range of the parameters τ, σ is allowed.

3.3. Consistent non-convex approach

In [30], a non-convex approach to sparsity-based audio declipping is proposed, called SPADE (sparse audio declipper). It is based on a formulation related to (10) with the sparse prior and declipping-type feasible set $\Gamma_\mathrm{T}^\mathrm{dec}$. However, it uses a different approach to the NP-hard minimization of the ℓ_0 pseudonorm. To provide at least a basic insight, the idea is to relax the strict relationship between the T and TF domains (governed deterministically by the transform) and utilize the iterative scheme of the alternating direction method of multipliers [31, 32]. The algorithm then consists of repetitive hard thresholding of the TF coefficients and the projection of a signal onto $\Gamma_\mathrm{T}^\mathrm{dec}$.

Following the idea of that paper, it is straightforward to build a more general algorithm than SPADE using the comprehensive set Γ_T defined in (6). Nonetheless, it is much more demanding to use

Algorithm 3: The CV algorithm for solving the general formulation (10), assuming the use of a tight frame

Input: The linear operators W, K, L, the proximal operator $\mathrm{prox}_{\mathcal{S}}$ and the projectors $\mathrm{proj}_{\Gamma_{\mathrm{T}}}, \mathrm{proj}_{\Gamma_{\mathrm{TF}}}$.

Choose the parameters $\tau, \sigma, \rho > 0$.

Choose the initial estimates $\mathbf{u}^{(0)}, \mathbf{v}_1^{(0)}, \mathbf{v}_2^{(0)}$.

for $i = 0, 1, \ldots$ **do**

/* update corresponding to h_1 */

$$\tilde{\mathbf{v}}_1^{(i+1)} = \mathbf{v}_1^{(i)} + \sigma W K \mathbf{u}^{(i)} - \sigma \mathrm{prox}_{\mathcal{S}/\sigma}\left(\mathbf{v}_1^{(i)}/\sigma + W K \mathbf{u}^{(i)}\right)$$

$$\mathbf{v}_1^{(i+1)} = \rho \tilde{\mathbf{v}}_1^{(i+1)} + (1-\rho)\mathbf{v}_1^{(i)}$$

/* update corresponding to h_2 */

$$\tilde{\mathbf{v}}_2^{(i+1)} = \mathbf{v}_2^{(i)} + \sigma K \mathbf{u}^{(i)} - \sigma \mathrm{proj}_{\Gamma_{\mathrm{TF}}}\left(\mathbf{v}_2^{(i)}/\sigma + K \mathbf{u}^{(i)}\right)$$

$$\mathbf{v}_2^{(i+1)} = \rho \tilde{\mathbf{v}}_2^{(i+1)} + (1-\rho)\mathbf{v}_2^{(i)}$$

/* notation for better readibility */

$$\mathbf{w} = \mathbf{u}^{(i)} - \tau K^* W^* \left(2\tilde{\mathbf{v}}_1^{(i+1)} - \mathbf{v}_1^{(i)}\right) - \tau K^* \left(2\tilde{\mathbf{v}}_2^{(i+1)} - \mathbf{v}_2^{(i)}\right)$$

/* update of $\mathbf{u}$ */

$$\tilde{\mathbf{u}}^{(i+1)} = \mathbf{w} + L^*\left(\mathrm{proj}_{\Gamma_{\mathrm{T}}}(L\mathbf{w}) - L\mathbf{w}\right)$$

$$\mathbf{u}^{(i+1)} = \rho \tilde{\mathbf{u}}^{(i+1)} + (1-\rho)\mathbf{u}^{(i)}$$

end

Output: $\mathbf{u}^{(i+1)}$

the non-convex approach for the universal formulation (10) and built the SPARE, i.e. the sparse audio restorer. The reason is that the projection onto $\Gamma_L \cap \Gamma_K$ is needed. Here, the Dykstra's projection algorithm [33] could be used, together with the projection lemma from [29], which is valid also for complex box-type sets, although this was not presented in therein. However, incorporating an iterative subroutine into an iterative algorithm is unfavorable.

3.4. Inconsistent convex approach

So far, the solutions to all of the restoration tasks were assumed to be consistent with the observed signal (or TF coefficients, or both). However, this assumption may be too strong, for example in the case of noisy data. In such a case, instead of strictly forcing the signal to lie in Γ_{T} and the coefficients to lie in Γ_{TF}, we minimize the distances to these sets. Formulation (13) would cover also this case, had we used the distance from Γ_{T} and Γ_{TF} instead of the indicator functions (which force the distance to be zero).

Since the proximal operator of a distance function of a point from a closed convex set is available [25], the inconsistent problem could be solved by the CV algorithm, similarly to the consistent one in Sec. 3.1 or 3.2.

3.5. Inconsistent non-convex approach

Similarly to the previous approach, the inconsistent non-convex approach is naturally obtained by modifying the consistent one. As mentioned above, the consistent SPARE algorithm would involve a projection onto $\Gamma_L \cap \Gamma_K$ in each iteration, ensuring the consistency of the resulting signal. Relaxing this step such that it corresponds to the proximal operator of distance from the set $\Gamma_L \cap \Gamma_K$ directly produces the inconsistent variant of SPARE.

4. EXPERIMENT

We perform an experiment that serves as the proof of concept of the presented general recovery formulation. Nevertheless, on top of that, the results suggest interesting implications that could lead to new consequences in audio coding.

4.1. Design of the experiment

The task is to restore a signal where some samples are missing; the present samples are moreover quantized. At the same time, a partial and quantized observation of the TF coefficients of the original (non-distorted) signal is provided. See Fig. 1 for an example. The goal is to illustrate that it is beneficial to utilize the double-domain approach compared to the restoration using only information in the time domain (abbreviated as T domain in some of the figures). The relaxed sparse prior, i.e. the ℓ_1 norm is used, hence leading to the consistent convex approach from Sec. 3.2.

The percentage of available samples/coefficients varies from $10\,\%$ up to $90\,\%$. In the time domain, the reliable samples are distributed (uniformly) randomly. In the TF domain, the coefficients largest in magnitude are kept (see Sec. 4.3 for additional comments on the choice of the coefficients). The uniform quantization is done by limiting the number of bits per sample or coefficient. For a given bit depth B (i.e. the number of bits used for representing each number, bps), $\Delta = 2^{-B+1}$ denotes the distance of two consecutive quantization levels. The quantized observation u^{q} of a real value u, $-1 \leq u \leq 1$ is computed using the so-called *mid-riser uniform quantizer* [23] as

$$u^{\mathrm{q}} = \mathrm{sgn}^{+}(u)\left(\left\lfloor \frac{|u|}{\Delta} \right\rfloor + \frac{1}{2}\right), \tag{21}$$

where $\mathrm{sgn}^{+}(u)$ returns 1 for $u \geq 0$ and -1 for $u < 0$. The bit depths are chosen as the powers of two, $B \in \{2, 4, 8, 16, 32\}$.

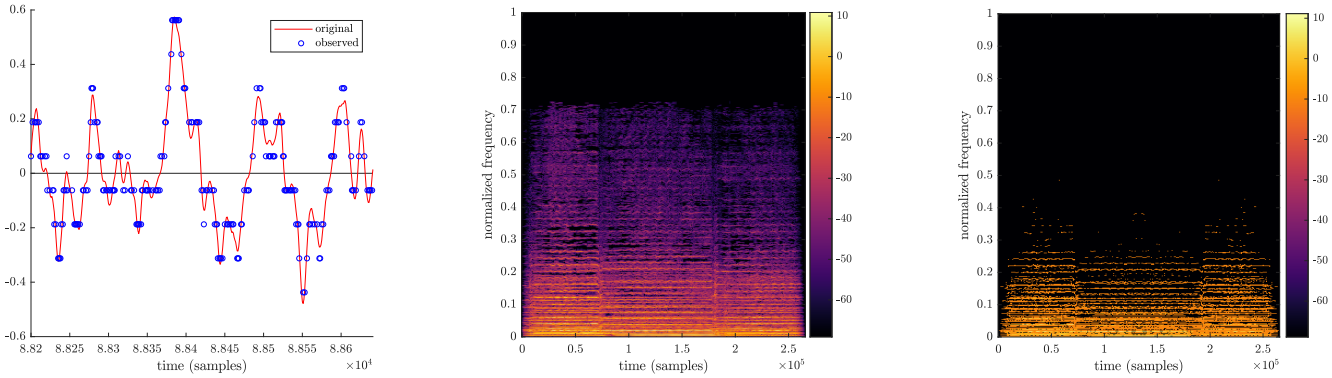
As the TF transform, the discrete Gabor transform is used, with sine window of length 2048 samples, 50% overlap and 2048 frequency channels. Such a transform produces a twice redundant tight frame, which is then normalized to obtain a Parseval tight frame. As the penalty, $\mathcal{S} = \|\cdot\|_1$ is used with no weighting, i.e. $W = \mathrm{Id}$. In order to evaluate the results, the PEMO-Q ODG score [34] and the SDR are measured, the latter being defined as

$$\mathrm{SDR}(\mathbf{y}, \hat{\mathbf{y}}) = 10 \log_{10} \frac{\|\mathbf{y}\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}, \tag{22}$$

where $\mathbf{y}$ is the original (non-distorted) time-domain signal and $\hat{\mathbf{y}}$ is the reconstruction. The result is expressed in decibels.

The experiment is run for a set of audio signals of varying complexity, from a single instrument up to a musical group. The signals are sampled at 44.1 kHz and they are originally ca 5 s long. However, to reduce the computational time, the proof-of-concept experiment only uses one-second long excerpts. For the purpose of quantization, these excerpts are also peak-normalized such that the maximum absolute value of each signal equals one.

Note that the results are visualized only for a single audio signal, `group_of_four`. For the rest of the results, see the link in Sec. 4.4.



(a) piece of the observed quantized signal (b) spectrogram of the original signal (c) spectrogram with kept quantized values

Figure 1: Data available to the decoder: (a) subsampled and quantized time-domain samples, and (c) subsampled and quantized TF coefficients. Although the real and imaginary parts are treated separately in the TF domain, the magnitude spectrogram is depicted here.

4.2. Results

4.2.1. Comparison with fixed bit depth

In the first visualization in Figure 2, the bit depth is fixed. The result corresponding to the single-domain approach with a given fraction of reliable samples in time domain serves as the reference (here, reliable means quantized but not completely unknown). These two parameters—bit depth and fraction—define the total bit rate of reliable information used in the single-domain approach.

This reference scenario is compared to different distributions of the total amount of bits between the time and TF domains while using the previously fixed bit depth. Note that only a limited number of options how to distribute the information between the time and the TF domains was tested.

Figure 2 shows the results for a single audio excerpt. Both evaluation metrics (ODG and SDR) are depicted. For bit depth 4, we present the results using both the analysis and the synthesis models (plots 2a, 2b, 2d, 2e). Since no significant difference between the performance of the analysis and the synthesis approaches is observed, only the analysis model is used for further comparison with the performance using bit depth 16 (plots 2c and 2f).

4.2.2. Comparison with variable bit depth

In the visualization in Figure 3, the number of bits per sample or coefficient varies. Two ways to display the results are used.

The single-domain approach is represented by the colored equibital lines[2]. The line color represents the restoration quality, according to the side colorbar. The line width represents the bit depth and the position represents the bit rate (in this case, only time-domain information is used).

The double-domain approach is represented by the colored points. Once again, the color indicates the restoration quality, the point size represents the bit depth (which is never different in the time and TF domains for a particular realization). Finally, the position represents the distribution of reliable information between the domains. Both in the case of lines and points, the following rule is applied: If more realizations with the same bit distribution appear, only the best of them is plotted.

[2]i.e., lines connecting points with the same total bit rate

The gray lines are also equibital. However, the total bit rate they represent is higher than what has been considered in the single-domain approach; thus no color can be assigned to these lines.

4.2.3. Discussion on the results

For a fixed number of bits per sample or coefficient, it is in general not beneficial to split the available information between the two domains; see the decrease in both ODG and SDR in the plots 2a, 2b, 2d and 2e with increasing percentage of reliable TF coefficients. The significant observation here is that the TF domain (in our setup) is able to provide useful information only with high bit depth—compare for example plots 2e and 2f.

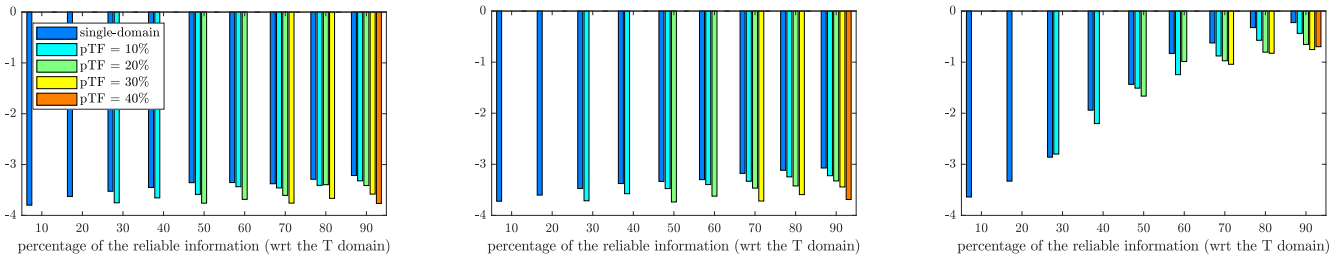
However, the scatter plots in Figure 3 show that there is a number of cases where it is useful to decrease the precision of the reliable samples and assign a part of the bit budget to the TF domain. Such a conclusion can be deduced from points which lie at an equibital line. If the color of the point indicates that the restoration quality is higher compared to the one indicated by the color of the line, it means that using the information in the TF domain instead of the time domain is beneficial.

4.3. On the choice and quantization of the TF coefficients

In the experiment, a tight Gabor frame is used to compute the TF representation of a real signal. Coefficients obtained using a Gabor frame actually attain a specific complex-conjugate structure.

In fact, only half of all the coefficients is needed; the other half may be computed as conjugate to the first half. Such a structure introduces a kind of redundancy, meaning that a pair of complex-conjugate coefficients contributes to the total bit rate by the same amount as a pair of real samples of the signal. This property is used in the implementation when choosing the set of reliable TF coefficients; it is ensured that for a given number of the reliable samples or coefficients, information from the TF domain yields the same bit rate as information from the time domain.

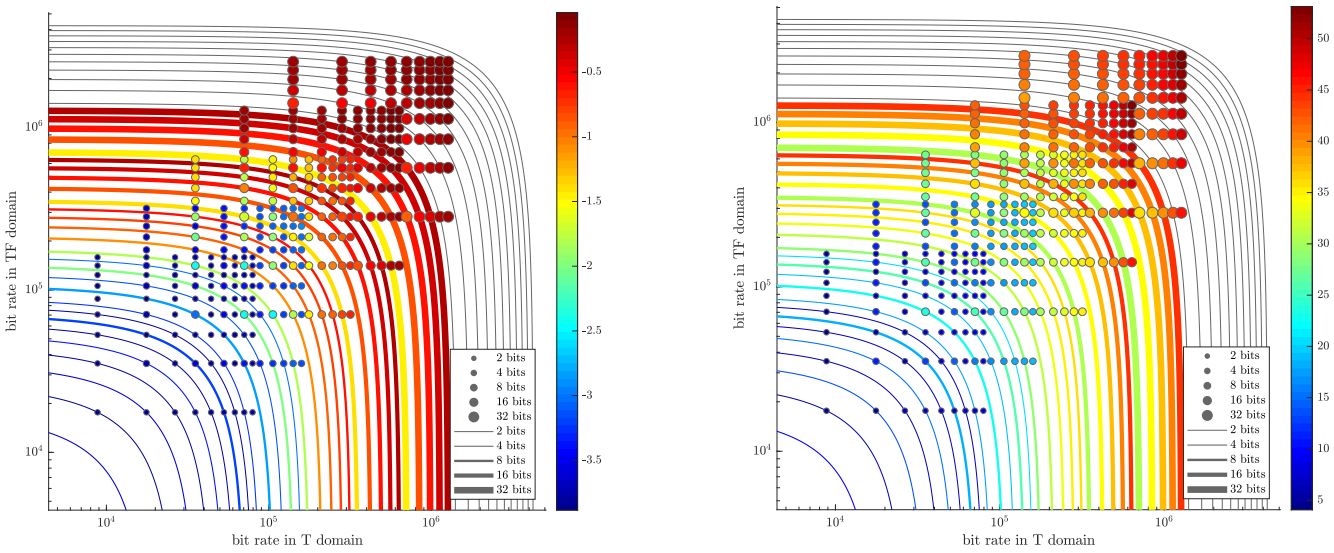
Furthermore, recall that the quantization defined by Eq. (21) is tailored for values from the interval $[-1, 1]$. To simulate the quantization for the observed TF coefficients $\mathbf{c}$, the quantization step Δ and all the quantization and decision levels in the TF domain are scaled by the factor of $\max\{\max\{|\Re(\mathbf{c})|\},\ \max\{|\Im(\mathbf{c})|\}\}$.



(a) ODG, synthesis model, 4 bits (b) ODG, analysis model, 4 bits (c) ODG, analysis model, 16 bits

(d) SDR, synthesis model, 4 bits (e) SDR, analysis model, 4 bits (f) SDR, analysis model, 16 bits

Figure 2: Bar plots. The legend shown in the first plot is common for all the plots; pTF denotes the percentage of reliable TF coefficients.



(a) ODG, analysis model (b) SDR, analysis model

Figure 3: Scatter plots.

4.4. Software and reproducible research

The experiment was run in MATLAB R2019b, using LTFAT [35] version 2.3.1. All the MATLAB codes, together with supplemental figures, are provided at `https://github.com/ondrejmokry/AudioRestorationFramework/`.

5. CONCLUSION

The paper provided a general flexible formulation not only covering multiple audio restoration tasks, but also allowing several degradation types to happen simultaneously. Another novelty is that the restoration can possibly take into account constraints in the time-frequency domain. The concept can be actually easily extended such that the reliable information is distributed among more than two different transform domains.

The aim of the experiment was not to outperform state-of-the-art methods in the field of audio restoration, but to show an application of the general formulation in a meaningful scenario. The general model is shown to be flexible enough to cover a rather complicated model of signal distortion, which included both dropouts and quantization of both the samples in the time domain and the TF coefficients. Even such a brief example demonstrates that it is worth studying possible distributions of reliable information.

6. ACKNOWLEDGMENTS

The work was supported by the Czech Science Foundation (GAČR) project number 20-29009S.

7. REFERENCES

[1] L. Rencker et al., "Sparse recovery and dictionary learning from nonlinear compressive measurements," *IEEE Trans. Signal Processing*, 2019.

[2] C. Gaultier et al., "A modeling and algorithmic framework for (non)social (co)sparse audio restoration," 2017.

[3] C. Gaultier, *Design and evaluation of sparse models and algorithms for audio inverse problems*, Theses, Université Rennes 1, 2019.

[4] Ç. Bilen, A. Ozerov, and P. Pérez, "Solving time-domain audio inverse problems using nonnegative tensor factorization," *IEEE Trans. Signal Processing*, 2018.

[5] A. Adler et al., "Audio Inpainting," *IEEE Trans. Audio, Speech, and Language Processing*, 2012.

[6] K. Siedenburg, M. Kowalski, and M. Dorfler, "Audio declipping with social sparsity," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014.

[7] O. Mokrý and P. Rajmic, "Reweighted l1 minimization for audio inpainting," in *Proc. of the SPARS workshop*, 2019.

[8] P. Záviška, P. Rajmic, and J. Schimmel, "Psychoacoustically motivated audio declipping based on weighted l1 minimization," in *International Conference on Telecommunications and Signal Processing*. IEEE, 2019.

[9] S. Shlien, "Guide to MPEG-1 audio standard," *IEEE Trans. Broadcasting*, 1994.

[10] L. Jacques, D. K. Hammond, and J. M. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine," *IEEE Trans. Information Theory*, 2011.

[11] P. Peter, J. Contelly, and J. Weickert, "Compressing audio signals with inpainting-based sparsification," in *Scale Space and Variational Methods in Computer Vision*, 2019, Springer.

[12] K. Gröchenig, *Foundations of time-frequency analysis*, Birkhäuser, 2001.

[13] P. Balazs et al., *Frame Theory for Signal Processing in Psychoacoustics*, Springer, 2017.

[14] P. Záviška et al., "Revisiting synthesis model in sparse audio declipper," in *Latent Variable Analysis and Signal Separation*. Springer, 2018.

[15] O. Derrien, T. Necciari, and P. Balazs, "A quasi-orthogonal, invertible, and perceptually relevant time-frequency transform for audio coding," in *European Signal Processing Conference*. IEEE, 2015.

[16] O. Christensen, *Frames and Bases, An Introductory Course*, Birkhäuser, 2008.

[17] A. J. E. M. Janssen, R. N. J. Veldhuis, and L. B. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Trans. Acoustics, Speech and Signal Processing*, 1986.

[18] W. Etter, "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Trans. Signal Processing*, 1996.

[19] M. Lagrange, S. Marchand, and J.-B. Rault, "Long interpolation of audio signals using linear prediction in sinusoidal modeling," *J. Audio Eng. Soc*, 2005.

[20] O. Mokrý et al., "Introducing SPAIN (SParse Audio INpainter)," in *European Signal Processing Conference*. IEEE, 2019.

[21] C.-T. Tan, B. C. J. Moore, and N. Zacharov, "The effect of nonlinear distortion on the perceived quality of music and speech signals," *J. Audio Eng. Soc*, 2003.

[22] J. Málek, "Blind compensation of memoryless nonlinear distortions in sparse signals," in *European Signal Processing Conference*. IEEE, 2013.

[23] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Information Theory*, 1998.

[24] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. of The National Academy of Sciences*, 2003.

[25] P. L. Combettes and J. C. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, 2011.

[26] L. Condat, "A generic proximal algorithm for convex optimization—application to total variation minimization," *Signal Processing Letters*. IEEE, 2014.

[27] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Advances in Computational Mathematics*, 2011.

[28] J. J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la société mathématique de France*, 1965.

[29] P. Rajmic et al., "A new generalized projection and its application to acceleration of audio declipping," *Axioms*, 2019.

[30] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: a flexible non-convex approach," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015.

[31] S. P. Boyd et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers.," *Foundations and Trends in Machine Learning*, 2011.

[32] P. Záviška, O. Mokrý, and P. Rajmic, "S-SPADE Done Right: Detailed Study of the Sparse Audio Declipper Algorithms," 2018, URL: `https://arxiv.org/pdf/1809.09847.pdf`.

[33] J. P. Boyle and R. L. Dykstra, "A method for finding projections onto the intersection of convex sets in hilbert spaces," in *Advances in Order Restricted Statistical Inference*. Springer, 1986.

[34] 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 Language Processing*, 2006.

[35] Z. Průša et al., "The Large Time-Frequency Analysis Toolbox 2.0," in *Sound, Music, and Motion*. Springer, 2014.