BRNO UNIVERSITY OF TECHNOLOGY

FACULTY OF ELECTRICAL ENGINEERING
AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

Ing. Pavel Záviška

AUDIO SIGNAL DECLIPPING AND DEQUANTIZATION USING SPARSITY-BASED METHODS

DECLIPPING A DEKVANTIZACE AUDIO SIGNÁLŮ POMOCÍ METOD ZALOŽENÝCH NA ŘÍDKÝCH REPREZENTACÍCH

SHORTENED VERSION OF PH.D. THESIS

Specialization:	Teleinformatics
Supervisor:	prof. Mgr. Pavel Rajmic, Ph.D.
Opponents:	prof. Ing. Zbyněk Koldovský, Ph.D. doc. Ing. Filip Šroubek, Ph.D., DSc.
Date of Defense:	

KEYWORDS

Audio, clipping, quantization, declipping, dequantization, restoration, sparsity, inverse problems, optimization, psychoacoustics.

KLÍČOVÁ SLOVA

Audio, clipping, kvantizace, declipping, dekvantizace, restaurování, řídkost, inverzní problémy, optimalizace, psychoakustika.

ARCHIVED IN

Dissertation is available at the Science Department of Dean's Office FEEC, Brno University of Technology, Technická 10, Brno, 616 00

MÍSTO ULOŽENÍ PRÁCE

Disertační práce je k dispozici na Vědeckém oddělení děkanátu FEKT VUT v Brně, Technická 10, Brno, 616 00

ISBN 80-214-

ISSN 1213-4198

CONTENTS

INTRODUCTION

Sound and music has always been a significant part of human culture, resulting in the natural need of recording and storing the music information. However, and audio signals are by nature disposed to different types of quality degradation. The degradations may arise directly during the recording process, they can be caused by the damage of the medium (such as the wax cylinder, LP, CD, etc.) or they can occur during transmission or streaming of the audio file.

There are many types of signal corruption. One of the most common is *noise*, which is usually described as an interference of the useful signal with an undesired signal that carries no useful information. Another very common type of signal degradation is *clipping*, which causes limitation of the dynamic range and thus loss of information in the peaks of the signal. The loss of samples can also be considered as a type of audio signal degradation.

Degradation of the signal does not necessarily need to be caused by accident. It can also be performed on purpose in order to reduce the size of the audio file. One may mention the quantization of the signal samples in the time domain or the lossy audio compression.

Typically, the corruption of the signal is irreversible and besides the perceptual quality of audio, it also affects several other fields such as automatic speech recognition in voice-controlled systems, medical diagnosis based on patient's speech analysis, compression and coding of audio signals in transmission systems, and many more. Therefore, to achieve a sufficient (or at least improved) perceptual quality, or to enhance the performance of systems that work with corrupted audio signals, it is necessary to perform restoration of the damaged audio signal.

The restoration tasks are usually formulated as inverse problems, handling each type of degradation individually; the restoration of the noisy signal is referred to as *denoising*, computing the missing samples is called *inpainting*, and recovery of the clipped or quantized samples is known as *declipping* and *dequantization*, respectively. Even though the restoration tasks can be approached in a similar way, each task is rather specific and requires satisfying different conditions based on the type of the restoration task. For this reason, the Thesis is mainly focused on clipping as one of the most common type of audio signal degradation and the corresponding restoration task—declipping. However, part of the Thesis is also devoted to the adaptation of declipping algorithms to the problem of audio dequantization.

Focusing purely on declipping, there are several commonly available tools that are able to find and repair clipped segments of audio signals. Nevertheless, the greatest weakness of these tools is that they are designed to be fast, simple, and user-friendly and thus they are usually based on interpolation, which is suitable only for the restoration of mildly clipped signals. In the case of moderate or severe clipping, these tools cannot fully remove the negative effects of clipping and may even produce artifacts that might degrade the perceived audio quality even more than the clipping itself. This further motivates scientists and audio engineers to develop new audio declipping methods that are able to deliver the best possible restoration quality with the lowest possible computational complexity. This work builds on previous research in the field of audio restoration and aims at proposing and implementing effective audio restoration methods with the primary focus on audio declipping and dequantization.

1 CLIPPING AND QUANTIZATION

This chapter is devoted to the nonlinear damage of the audio signal this Thesis works with, i.e., clipping and quantization. The first section discusses clipping and explains the basic questions about clipping (what clipping is, where it may arise, why it causes problems, etc.), followed by a section, which discusses the general formulation of the declipping problem. Next follows the section on quantization, where different types of quantization are outlined and the quantization models used in this work are described and illustrated on examples. Similarly to declipping, the basic idea of the dequantization problem is specified in the final part of this chapter.

Clipping

Clipping can be described as a nonlinear form of signal distortion affecting peaks of the signal. It usually occurs when a signal exceeds its allowed dynamic range and the signal peaks get clipped to the boundaries of the dynamic range. Thus, information located in the peaks is lost. From the frequency-domain perspective, such a nonsmooth phenomenon naturally produces artificial higher harmonics. The newly-introduced higher harmonics shift the signal energy towards higher frequencies, which may cause trouble in some applications.

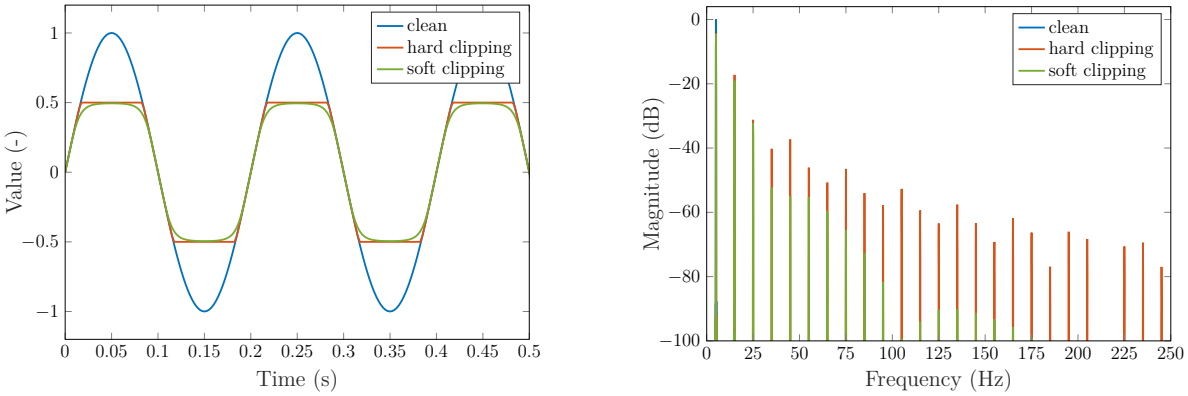
Even though clipping may affect any type of signal, the most common occurrence of this artifact is with audio signals where causes undesirable and perceptually unpleasant artifacts. Clipping may occur during the recording stage when the input gain on the recording equipment is set a bit too high. Also recording loud sounds using microphones with low dynamic range (typically integrated in a notebook or mobile phone) may result in clipping. In the analog domain, clipping is very often caused in amplifiers by the limited range of output transformers.

Clipping is an undesirable effect that may cause several problems. Not only has clipping a significant negative effect on the perceptual quality of the signal [1]. Several studies show that it also degrades the accuracy of automatic speech recognition [2, 3, 4], causes problems in LPC prediction, resulting in an inaccurate estimation of LPC [5], or degrades the accuracy of voice-based Parkinson's disease detection [6]. During reproduction, severe clipping can even damage the loudspeaker [7].

According to the character of clipping, two different types of clipping can be distinguished—*hard clipping* and *soft clipping*. The effect of both types is demonstrated on a sine wave both in the time-domain (see Fig. 1.1a) and in the magnitude spectrum (see Fig. 1.1b). In the case of hard clipping, samples of the signal $\mathbf{x} \in \mathbb{R}^N$ are limited to fit the dynamic range given by *clipping thresholds* $[-\theta_\mathrm{c}, \theta_\mathrm{c}]$. The clipped signal $\mathbf{y} \in \mathbb{R}^N$ can be formally prescribed as

$$y_n = \begin{cases} x_n & \text{for} \quad |x_n| < \theta_\mathrm{c}, \\ \theta_\mathrm{c} \cdot \mathrm{sgn}(x_n) & \text{for} \quad |x_n| \geq \theta_\mathrm{c}, \end{cases} \tag{1.1}$$

where the subscript n refers to the n-th sample of the signal, and sgn represents the signum function.



(a) waveform

(b) magnitude spectrum

Fig. 1.1: Demonstration of the hard clipping and soft clipping on a sine wave. The frequency of the sine wave is 5 Hz and the sampling frequency is 500 Hz. The clipping threshold for both hard clipping and soft clipping was set to $\theta_c = 0.5$. The magnitude spectra were generated from 20 seconds of audio, and the Blackman window was used to attenuate the side lobes of the spectra.

Declipping

By the term *declipping*, it is meant the inverse task of estimating the original signal $\mathbf{x}$ from the clipped observation $\mathbf{y}$. The goal of declipping is to provide signals most similar to the unknown original reference or at least to remove the disturbing phenomena caused by clipping.

In line with (1.1), the indexes of signal samples can be divided into three disjoint sets R, H and L, such that $R \cup H \cup L = \{1, \ldots, N\}$, which correspond to the positions of *reliable* (not influenced by clipping) samples, and samples that have been clipped to the high clipping threshold θ_c and low clipping threshold $-\theta_c$, respectively. To select only samples from the specific set, the respective restriction operators M_R, M_H and M_L (also called *masks*) are used. These operators can also be viewed as matrices, which are formed from the identity matrix $N \times N$ by removing the respective rows that do not belong to the selection.

In the declipping restoration task, it is natural to desire that the recovered signal $\hat{\mathbf{x}}$ should match the clipped signal $\mathbf{y}$ at the reliable positions, and at the clipped positions, its samples should lie above θ_c or below $-\theta_c$. Such conditions can be formalized by defining a (convex) set of time-domain signals Γ as follows:

$$\Gamma = \{\tilde{\mathbf{x}} \in \mathbb{R}^N \mid M_R\tilde{\mathbf{x}} = M_R\mathbf{y}, M_H\tilde{\mathbf{x}} \geq \theta_c, M_L\tilde{\mathbf{x}} \leq -\theta_c\}, \tag{1.2}$$

where the inequalities are considered elementwise. Such an approach, where the reconstructed signal $\hat{\mathbf{x}}$ is forced to lie in the set of feasible solutions Γ, i.e., $\hat{\mathbf{x}} \in \Gamma$, is called *consistent* or *fully consistent*. This approach is necessary to obtain the restored signal $\hat{\mathbf{x}}$ that is as close to the unknown original signal $\mathbf{x}$ as possible. On the other hand, it is possible to sometimes break the consistency of the solution and allow some deviation on reliable samples. Such solutions are referred to as *R-inconsistent*.

Quantization

Quantization is the mapping of continuous amplitude values to the nearest quantization levels that can be represented by a finite number of bits [8]. This step is inevitably lossy and a *quantization error* $\mathbf{e}$ is introduced, defined as the difference between the original and quantized signal. Formally, the general concept of quantization can be prescribed as

$$(y^{\mathrm{q}})_n = x_n - e_n, \tag{1.3}$$

where $\mathbf{y}^{\mathrm{q}}$ denotes the quantized signal, $\mathbf{x}$ is the original signal and $\mathbf{e}$ represents the quantization error. The Thesis focuses mainly on uniform mid-riser quantization, according to which the quantized signal $\mathbf{y}^{\mathrm{q}} \in \mathbb{R}^N$ is obtained as

$$(y^q)_n = \mathrm{sgn}^+(x_n) \cdot \Delta \cdot \left(\left\lfloor \frac{|x_n|}{\Delta} \right\rfloor + \frac{1}{2} \right), \tag{1.4}$$

where n denotes the n-th sample of the signal, Δ is the quantization step given by $\Delta = 2 \cdot \frac{1}{2^w}$, and sgn^+ denotes the altered signum function, returning 1 also for the zero input.

Dequantization

Dequantization, similarly to declipping is the inverse task of estimating the original signal $\mathbf{x}$ from its quantized observation $\mathbf{y}^{\mathrm{q}}$.

There is a number of reasons why dequantization is an important task and has its application. In some cases, where the original audio was recorded with low dynamic range or needs to be further edited, the standard 16 bps bit depth could be insufficient. Another application of dequantization may arise in special cases, where less than the standard bit depth has to be used. This can typically occur in communication systems due to bandwidth limitations [9, 10]. Recently, the need to enhance the bit-depth of audio signals appeared in artificial audio generation using a Flow-based Neural Vocoder [11].

From the definition of the uniform quantization, we can assume that the unknown original sample x_n lied no further than half of the quantization step from its current quantization level y_n^{q}. Thus, the searched unknown vector $\tilde{\mathbf{x}} \in \mathbb{R}^N$ should fulfill the following requirement:

$$\forall n \colon y_n^{\mathrm{q}} - \frac{\Delta}{2} \leq \tilde{x}_n \leq y_n^{\mathrm{q}} + \frac{\Delta}{2}. \tag{1.5}$$

The dequantization conditions can be formalized by defining the convex set Γ as follows:

$$\Gamma = \left\{ \tilde{\mathbf{x}} \mid \|\tilde{\mathbf{x}} - \mathbf{y}^{\mathrm{q}}\|_\infty \leq \frac{\Delta}{2} \right\}, \tag{1.6}$$

and then require the dequantized signal to lie in this set, formally $\tilde{\mathbf{x}} \in \Gamma$. As in the declipping case, strictly forcing $\tilde{\mathbf{x}}$ to lie in Γ is called the *consistent* approach but it is also possible to extend the allowed interval for each quantization level and thus allow some deviation from Γ.

2 THESIS AIMS AND OBJECTIVES

The main aim of the Thesis is to propose and implement effective methods and algorithms for the restoration of corrupted audio signals with the primary focus on audio declipping.

To do so, the declipping task will be first formulated as an optimization problem, and then optimization algorithms will be chosen to solve the problems. The developed methods can be further improved by involving additional information about the signal, such as psychoacoustic information or information concerning the characteristics of the clipped samples. A special focus is also paid to improving the results obtained by methods inconsistent in the reliable part.

A necessary part of the Thesis is the evaluation of the achieved results, which will be performed on a common dataset using several evaluation metrics. Selected algorithms will also be applied to the problem of audio dequantization and evaluated using the same metrics as in the case of declipping.

Following the idea of reproducible research, the implementations of the algorithms for audio declipping and dequantization will be made publicly available.

Formulation of the declipping problem

First, the declipping problem using sparse representations will be formulated. This task seems rather simple, but there are still several possibilities how the problem can be formulated. A critical role plays the sparsity promoting regularizer. It can be hard thresholding approximating the nonconvex ℓ_0-norm, soft thresholding being the proximal operator of the convex ℓ_1 norm or possibly a shrinkage operator promoting a structure of the time-frequency coefficients.

There are two possible approaches to signal modeling—the synthesis and analysis models. The Thesis will explore both signal models and compare them in different modeling schemes.

Also, the set of feasible solutions can be formulated in multiple ways. The main issue is whether the problem should always obey the full consistency according to (1.2) or whether a slight deviation on the reliable samples could bring an improved perceptual quality of the reconstructed signal.

Selecting the optimization algorithm

Since finding the ideal solution to the recovery problem is NP-hard in most of the cases, the solution is usually approximated and numerically solved using an optimization algorithm.

For convex optimization problems, the Thesis will focus primarily on proximal splitting methods. Nonconvex problems will be approached by the means of ADMM.

It is also possible to explore and experiment with different types of optimization algorithms. The aim is to find an algorithm with sufficient accuracy, fast convergence, robustness, and low computational expenses, although restoration quality remains the main goal.

Apart from delivering new algorithms, the aim of the Thesis is also to improve existing ones. For instance, a one-step projection could significantly speed up the synthesis model-based restoration tasks. Also, it was found out that in [12], the presented synthesis variant of the SPADE algorithm (S-SPADE) does not fit the ADMM paradigm. Therefore, finding a proper synthesis variant of SPADE is also one of the goals of the Thesis.

Adding a priori information

Even though methods purely based on a sparsity assumption can obtain good restoration results, there is still room for improvement. Considering some additional assumptions about the signal may significantly improve the perceived quality of restoration.

One of the promising ways is to involve psychoacoustics in the restoration task, which should help to restore mainly perceptually significant coefficients and thus improve the perceived restoration quality. Also, information about the distribution of spectral components introduced by clipping could be used to distinguish the original spectral components and the distortion components.

The Thesis will be looking for ways to obtain and implement the above-mentioned information into the restoration algorithms.

Replacing reliable samples

Some of the existing audio declipping algorithms produce solutions inconsistent in the reliable part with the option to force the consistency in the postprocessing step. Such a task naturally increases the SDR, however to the best of our knowledge, no study examined what effect this postprocessing replacement has on the perceived audio quality. Therefore, this part of the Thesis will study the results and consequences of the mentioned replacement. Also, an effort will be made to introduce novel methods for quality enhancement of the inconsistent declipping methods exploiting the knowledge of reliable samples.

Evaluation

An indispensable part of the Thesis is the evaluation of the obtained results from the implemented algorithms. The results of the methods included in this Thesis will be evaluated and compared to other state-of-the-art methods. To evaluate the quality of restoration, classical error measures such as the signal-to-distortion ratio (SDR), and perceptually motivated objective evaluators like PEAQ or PEMO-Q will be used.

A majority of previous research papers on audio declipping used various audio datasets, in most of the cases sampled at 16 kHz. Such a low sampling frequency has been used mainly for computational reasons. One of the goals of this Thesis is to compare existing audio declipping approaches on a common dataset with excerpts sampled at 44.1 kHz, i.e., the standard audio quality. This dataset, created specifically for this task, will be publicly available to enable the comparison of the declipping methods developed in the future with the already existing ones.

Audio dequantization

As indicated in Chapter 1, audio declipping and dequantization are very similar tasks, although audio dequantization has gained far less research interest than declipping. Therefore, as a part of the Thesis, the selected audio declipping algorithms will be adapted to solve the dequantization problem to examine whether successful audio declipping methods will also perform well in the dequantization case.

Algorithm implementation

Last but not least, GitHub repositories with MATLAB implementations of the developed declipping and dequantization algorithms will be created, containing also the testing audio excerpts. Moreover, a supplementary web page for audio declipping will be created, containing a comparison of different declipping methods with the option to compare the achieved results by listening to the declipped excerpts.

3 EXPERIMENT DESIGN AND EVALUATION

This chapter is devoted to a description of the experiments that will be performed and described later in this Thesis, in order to compare the achieved audio quality of the proposed restoration algorithms.

The audio dataset used for all the following experiments and evaluations consists of 10 musical excerpts in mono with an approximate duration of 7 seconds and a sampling frequency of 44.1 kHz. It was extracted from the database called "Sound Quality Assessment Material recordings for subjective tests" (SQAM)[1] provided by European Broadcasting Union (EBU). Selected audio tracks were transferred into mono signals by averaging the left and right channels, cut using the Adobe Audition CS6 to an approximate duration of 7 seconds (depending on the content of the excerpts), and saved as uncompressed WAV files with 16 bps bit depth and a sampling frequency of 44.1 kHz.

The clipped audio files were created by artificially hard clipping the input signals in agreement with the definition of hard clipping in Eq. (1.1). The amount of distortion added into the signals is quantified using the Signal-to-Distortion Ratio (SDR), which is for signals $\mathbf{u}$ and $\mathbf{v}$ defined as

$$\mathrm{SDR}(\mathbf{u}, \mathbf{v}) = 20 \log_{10} \frac{\|\mathbf{u}\|_2}{\|\mathbf{u} - \mathbf{v}\|_2}. \tag{3.1}$$

We performed several informal listening tests based on which we chose 7 different clipping levels, to cover the range from very harsh clipping to mild but still noticeable clipping. The selected input SDR values are 1, 3, 5, 7, 10, 15, and 20 dB.

For the dequantization experiments, we exploited the classical uniform mid-riser quantization according to Eq. (1.4) with word lengths ranging from 2 to 8 bps.

The quality of the restored signal $\hat{\mathbf{x}}$ is evaluated as $\mathrm{SDR}(\mathbf{x}, \hat{\mathbf{x}})$, where $\mathbf{x}$ represents the original signal and the SDR is computed using Eq. (3.1). We also define SDR_c, which is SDR computed only on the clipped samples, and improvement of the SDR denoted as ΔSDR. The similarity in waveforms may not necessarily imply perceptual quality, therefore, we also use perceptually-motivated measures, specifically PEAQ [13] and PEMO-Q [14]. Both metrics output the Objective Difference Grade (ODG), which measures the degradation of a test input relative to a reference input. and ranges from 0 (imperceptible degradation) to -4 (very annoying degradation).

Restoration algorithms based on sparse representations rely heavily on the representation used. Purely frequency transform, such as Discrete Fourier Transform (DFT), is not typically a good representative of the signal since audio signals are not stationary and the frequency changes over time. Therefore, we picked the Discrete Gabor Transform (DGT) as the time-frequency representation, with a Hann window as the used window function. For all experiments, the length of the window was set to 8,192 samples (approx. 186 ms) with 75 % overlap and 16,384 frequency channels.

[1] `https://tech.ebu.ch/publications/sqamcd`

4 AUDIO DECLIPPING ALGORITHMS

This chapter is devoted to a detailed description and comparison of various sparsity-based audio declipping algorithms, which forms one of the main contributions of this Thesis. It contains both the original algorithms developed by the author and the adopted algorithms, which, nevertheless, have been reimplemented or modified for better performance. Most of the algorithms have been published as a part of the audio declipping survey [15].

Consistent ℓ_1 minimization

First, we approach the declipping problem using the synthesis variant of ℓ_1-relaxed declipping problem, which in the uncoinstrained form takes form of

$$\arg\min_{\mathbf{z}} \|\mathbf{z}\|_1 + \iota_{\Gamma^*}(\mathbf{z}), \tag{4.1}$$

where $\mathbf{z} \in \mathbb{C}^P$ denotes signal coefficients in the transformed domain. To solve this problem, it is possible to use the Douglas–Rachford (DR) algorithm since the problem takes the form of a sum of two convex functions. The two main steps of the algorithm are soft thresholding as the proximal operator of ℓ_1-norm, and the projection onto the set Γ^* as the proximal operator of the respective indicator function ι_{Γ^*}, which for Parseval tight frames ($DD^* = Id$) and a box-type set Γ can be computed using the following closed-form formula:

$$\mathrm{proj}_{\Gamma^*}(\mathbf{z}) = \mathbf{z} - D^* \left(D\mathbf{z} - \mathrm{proj}_{\Gamma}(D\mathbf{z})\right), \tag{4.2}$$

where the inner projection step is a projection onto a box-type set and in the particular case of declipping can be computed as a simple time domain elementwise mapping

$$\Big(\mathrm{proj}_{\Gamma}(\mathbf{x})\Big)_n = \begin{cases} y_n & \text{for } n \in R, \\ \max(\theta_\mathrm{c}, x_n) & \text{for } n \in H, \\ \min(-\theta_\mathrm{c}, x_n) & \text{for } n \in L. \end{cases} \tag{4.3}$$

Before the explicit projector was developed, the projection had to be computed for all three sets R, H, and L separately, corresponding to the problem

$$\arg\min_{\mathbf{z}} \|\mathbf{z}\|_1 + \iota_{R^*}(\mathbf{z}) + \iota_H(D\mathbf{z}) + \iota_L(D\mathbf{z}), \tag{4.4}$$

where R^* denotes the set corresponding to the reliable samples in the transformed domain, formally

$$R^* = \{\tilde{\mathbf{z}} \in \mathbb{C}^P \mid M_\mathrm{R} D\tilde{\mathbf{z}} = M_\mathrm{R}\mathbf{y}\}. \tag{4.5}$$

The problem (4.4) can be optimized using the Condat–Vũ (CV) algorithm. The experiments conducted in the Thesis show that both algorithms converge to the same solution, however, the DR algorithm converges significantly faster making it the preferred choice.

The analysis variant of the consistent ℓ_1 relaxation problem is formulated as

$$\arg\min_{\mathbf{x}} \|A\mathbf{x}\|_1 + \iota_\Gamma(\mathbf{x}). \tag{4.6}$$

To solve this problem, the Chambolle–Pock (CP) algoriothm is used. Two principal steps of the algorithm are the clip function and projection onto Γ. The projection is computed in the time-domain, which is the same simple elementwise mapping as in (4.3), and the clip function is the result of Fenchel–Rockafellar conjugate of the soft thresholding as the proximal operator of the ℓ_1-norm.

The comparison of the DR and CP algorithms presented in the Thesis reveals that the synthesis approach tends to converge faster and produces consistently better results in terms of ΔSDR than its analysis variant using the Chambolle–Pock algorithm.

Reweighted ℓ_1 minimization

The idea of reweighting applied to audio declipping was published by Weinstein and Wakin [16] under the acronym Rℓ_1CC (Reweighted ℓ_1 with Clipping Constraints), and it was shown that reweighting can significantly improve the overall declipping performance. Nevertheless, the authors assume only the synthesis model of the signal and provide no algorithm to solve the optimization problem. The weighted variant of the problems read

$$\arg\min_{\mathbf{z}} \|\mathbf{w} \odot \mathbf{z}\|_1 + \iota_{\Gamma^*}(\mathbf{z}), \tag{4.7a}$$

$$\arg\min_{\mathbf{x}} \|\mathbf{w} \odot A\mathbf{x}\|_1 + \iota_\Gamma(\mathbf{x}), \tag{4.7b}$$

and the weights are computed from the current temporary solution $\mathbf{z}$ as $\mathbf{w} = \frac{1}{|\mathbf{z}|+\epsilon}$. The problems are solved via the DR algorithm and the CP algorithm for the synthesis variant (4.7a) and analysis variant (4.7b), respectively.

The experiments conducted in the Thesis show that in the synthesis case, the reweighting helps to improve the restoration quality according to the ΔSDR. However, a significant improvement can be observed only for the first three outer iterations. Then the performance in terms of ΔSDR levels out and even drops a little after reaching 8 iterations. On the other hand, the experiments suggest the dominance of the analysis approach, since the results improve with every outer iteration and outperform the synthesis variant.

R-inconsistent ℓ_1 minimization

For a long time, the approach proposed by Defraene *et al.* [17] was the only one to include psychoacoustics in declipping (both in the model itself and in the evaluation). The optimization task is based on the weighted ℓ_1-norm of the coefficients, it allows a deviation on the reliable samples and it is formulated as

$$\arg\min_{\mathbf{z}} \left\{ \frac{1}{2\lambda} \|M_\mathrm{R} D\mathbf{z} - M_\mathrm{R}\mathbf{y}\|_2^2 + \|\mathbf{w} \odot \mathbf{z}\|_1 \right\} \quad \text{s.t.} \quad D\mathbf{z} \in \Gamma_\mathrm{H} \cap \Gamma_\mathrm{L}. \tag{4.8}$$

To be more specific about the method, the signal is processed window-by-window, and the task (4.8) is solved independently for the signal chunks given by windowing. The optimization core of the algorithm (called CSL1) in the original paper [17] was built upon the CVX toolbox [18] but no implementation is available. Here, we decided to solve the optimization problem using a proximal algorithm, specifically the Condat–Vũ algorithm.

Social Sparsity

Siedenburg *et al.* [19] utilized the concept of social sparsity. The algorithm is based on solving the following optimization problem:

$$\min_{\mathbf{z}} \Big\{ \frac{1}{2}\|M_{\mathrm{R}} D\mathbf{z} - M_{\mathrm{R}}\mathbf{y}\|_2^2 + \frac{1}{2}\|h(M_{\mathrm{H}} D\mathbf{z} - M_{\mathrm{H}}\theta_{\mathrm{c}}\mathbf{1})\|_2^2 + \\ + \frac{1}{2}\|h(-M_{\mathrm{L}} D\mathbf{z} - M_{\mathrm{L}}\theta_{\mathrm{c}}\mathbf{1})\|_2^2 + \lambda\mathcal{R}(\mathbf{z}) \Big\}, \quad (4.9)$$

where the symbol $\mathbf{1}$ represents the vector of ones, which is as long as the signal. The deviation of the clipped samples from the feasible sets Γ_{H} and Γ_{L} is penalized using the *hinge* function h. Furthermore, $\mathcal{R}$ represents the regularizer of the time-frequency (TF) coefficients. The original paper [19] suggests using four types of shrinkage operators—LASSO (L), Windowed Group LASSO (WGL), Empirical Wiener (EW), and Persistent Empirical Wiener (PEW). The problem is numerically solved by the Iterative Shrinkage/Thresholding Algoritm (ISTA).

In the case of the social shrinkage operators WGL and PEW, it is necessary to specify the size of the coefficient neighborhood in the TF plane. For the test case of the experiments, the best-performing size of the neighborhood was 3×7 (i.e., 3 coefficients in the direction of frequency and 7 coefficients in time, symmetrically distributed around the point tf), which will be used further in the Thesis.

The Thesis also studies the convergence speed of the algorithm and exploits different strategies (ISTA, FISTA, Adaptive Restart (AR), and AR with threshold) to accelerate the convergence of the algorithm.

Consistent ℓ_0 approximation

Another successful approach to audio declipping called Sparse Audio Declipper (SPADE) was presented in [12], where the optimization problem is formulated using the ℓ_0-norm as

$$\arg\min_{\mathbf{x},\mathbf{z}} \|\mathbf{z}\|_0 \quad \text{s.t.} \quad \mathbf{x} \in \Gamma \text{ and } \|A\mathbf{x} - \mathbf{z}\|_2 \leq \varepsilon, \quad (4.10a)$$

$$\arg\min_{\mathbf{x},\mathbf{z}} \|\mathbf{z}\|_0 \quad \text{s.t.} \quad \mathbf{x} \in \Gamma \text{ and } \|\mathbf{x} - D\mathbf{z}\|_2 \leq \varepsilon, \quad (4.10b)$$

where (4.10a) and (4.10b) represent the problem formulation for the analysis and the synthesis variant, respectively, and ε is a selected parameter.

The signal is cut into overlapping blocks and windowed prior to processing. Therefore, in (4.10), $\mathbf{y}$ should be understood as one (and each) of the signal chunks. The overall resulting signal is made up by the overlap–add procedure. As the transform, SPADE algorithms use the (overcomplete) DFT.

This Thesis provides a basic derivation of the algorithms. For more details, we refer the reader to the report [20] The Thesis also describes exploiting the projection (4.2) to accelerate the original synthesis variant S-SPADE. Later, it was found out that the S-SPADE is not quite a synthesis counterpart of the A-SPADE because both optimization subtasks are carried over $\mathbf{z}$ (in the domain of coefficients). It can be shown that the problem formulation corresponding to the S-SPADE algorithm is

$$\min_{\mathbf{w},\mathbf{z}} \|\mathbf{z}\|_0 \quad \text{s.t.} \quad D\mathbf{w} \in \Gamma \ \text{ and } \ \|\mathbf{w}-\mathbf{z}\|_2 \leq \varepsilon. \tag{4.11}$$

Therefore, we developed a new synthesis variant of the SPADE algorithm, which is truly the synthesis counterpart of A-SPADE and solves (4.10b). This new algorithm is referred to as S-SPADE "Done Properly" and was published in [21].

Results and discussion

This section is designed to perform the overall comparison of the algorithms presented in this chapter, which are also compared with four additional algorithms—Constrained Orthogonal Matching Pursuit (C-OMP) [22], Dictionary Learning (DL) [23], Nonnegative Matrix Factorization (NMF) [24], and Janssen's method [25].

The comparison of the methods is in the Thesis performed using $\Delta\mathrm{SDR}_\mathrm{c}$, PEAQ, and PEMO-Q. The average results of the latter are displayed in Fig. 4.1.

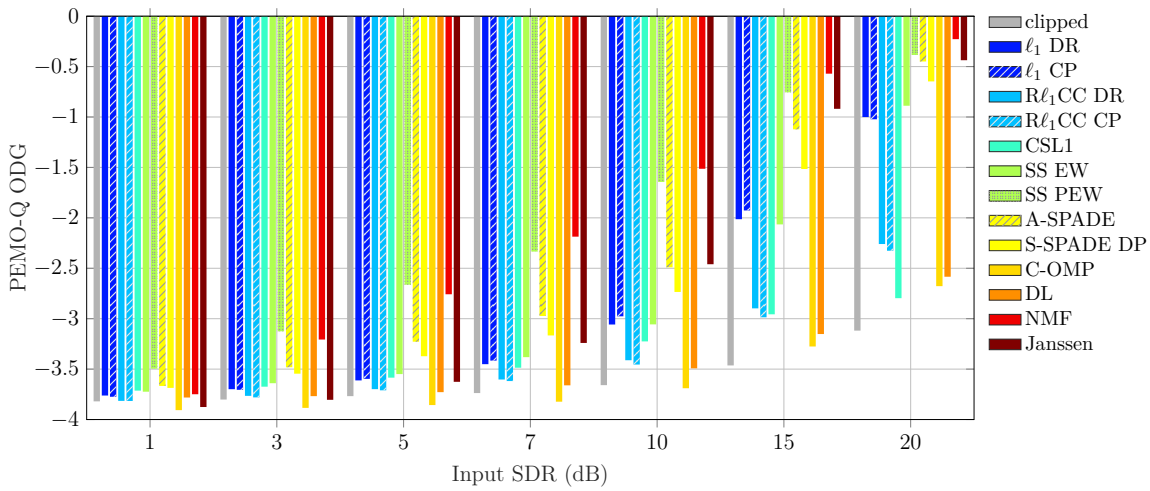


Fig. 4.1: Average declipping performance in terms of PEMO-Q ODG.

Apart from the restoration quality, the computational complexity of the algorithms is also evaluated in the Thesis. The average worst-case computational time per a second of audio can vary from 20 seconds for consistent ℓ_1 minimization to almost 1 hour for NMF.

5 INCORPORATING PSYCHOACOUSTICS INTO AUDIO DECLIPPING

The only algorithm exploiting any additional information based on psychoacoustics was by Defraene *et al.* [17], coined PCSL1. The authors utilized the effect of simultaneous masking and used the MPEG psychoacoustic model to weight the time-frequency coefficients during the ℓ_1-minimization, solving the problem (4.8).

In this chapter, we utilize the fully consistent optimization problems based on the weighted ℓ_1-minimization, are solved using the DR and CP algorithms for the synthesis and analysis variant, respectively. Moreover, three possible constructions of the weights are presented—based on the absolute threshold of hearing (ATH), on the global masking threshold (GMT), and on a quadratic curve. The presented approaches for the synthesis variant using the DR algorithm were published in a conference paper [26].

Absolute threshold of hearing

ATH represents a good indicator of the sensitivity of human ear at certain frequencies. Therefore, the main idea of using the ATH in the declipping task aims at eliminating the negative effects of clipping especially at frequencies where the human ear is most sensitive. This can be achieved by weighting the TF coefficients such that large weights correspond to frequencies with low respective ATH values and vice versa. Since the task of creating the vector of weights from the ATH is not straightforward, we examine the following three possibilities:

$$\mathbf{w}_{\mathrm{ATH1}} = (\mathbf{t} - \min(\mathbf{t}) + 1)^{-1}, \tag{5.1a}$$

$$\mathbf{w}_{\mathrm{ATH2}} = -\mathbf{t} + \tau, \tag{5.1b}$$

$$\mathbf{w}_{\mathrm{ATH3}} = 2 \cdot 10^{-5} \cdot 10^{(-\mathbf{t}+\tau)/20}, \tag{5.1c}$$

where $\mathbf{t}$ represents the vector of the ATH values for equispaced frequencies and τ is the parameter that sets the maximum value of the ATH in dB.

Global masking threshold

The information contained in the GMT can be used to focus on perceptually important components of the signal, while less audible components can be tolerated to a greater extent because they will be masked and thus not perceived. Consequently, the weights should be constructed in a similar way to the case of ATH, i.e., low values of GMT should produce large weights and vice versa. To do so, we utilized the same possibilities (5.1), only $\mathbf{t}$ now represents the GMT.

To compute the GMT from the obtained data, a slightly modified MPEG-1 Psychoacoustic Model 1 is used. The official standard is strictly limited to 512-sample long windows, and the used representation works with 8,192 samples long windows with 16,384 frequency channels. Hence, we used a slightly modified and simplified version of the psychoacoustic model, which is not restricted in terms of the block length.

Parabola-based weights

Apart from the ATH and GMT based variants, we also include a third option which is based on the idea that most of the energy in audio signals is usually concentrated at lower frequencies and that clipping introduces artificial higher harmonics that were not present in the original signal. Consequently, the weights are constructed in such a way that the higher harmonics are suppressed, while the lower frequencies are preserved.

A simple and effective approach to addressing this issue is to weigh the coefficients linearly, however, better restoration results are obtained when a second-order polynomial is used. Formally, these weights are for the real-valued DGT obtained as $\mathbf{w}_{\mathrm{p}} = \mathbf{m} \odot \mathbf{m}$, where $\mathbf{m} = [1, \ldots, \left\lfloor \frac{M}{2} \right\rfloor + 1]$, where M is the number of frequency bins of the DGT.

Results and discussion

The experiments conducted in the Thesis show that weighting with the GMT is a better idea than using a simple ATH curve. Also, the best variant of converting the GMT or ATH curves into the actual vector of weights $\mathbf{w}$ seems to be the one using the inversion, i.e., Eq. (5.1a). Nevertheless, among all the choices, the best results by far are obtained by the parabola weights, which produce approximately 10 dB better results than the plain ℓ_1 minimization.

The overall PEMO-Q results are illustrated in Fig. 5.1. In this comparison, we include all three weight types (ATH, GMT, parabola) along with the nonweighted variants. For comparison, we include Defraene's approaches [17] CSL1 (nonweighted variant), PCSL1 (weighted using the GMT), and we also incorporate the parabolic weights into this algorithm, leading to a Parabola-weighted CSL1 (PWCSL1). For reference, the two best-performing algorithms from Chapter 4, i.e., SS PEW and NMF, are also part of the evaluation.

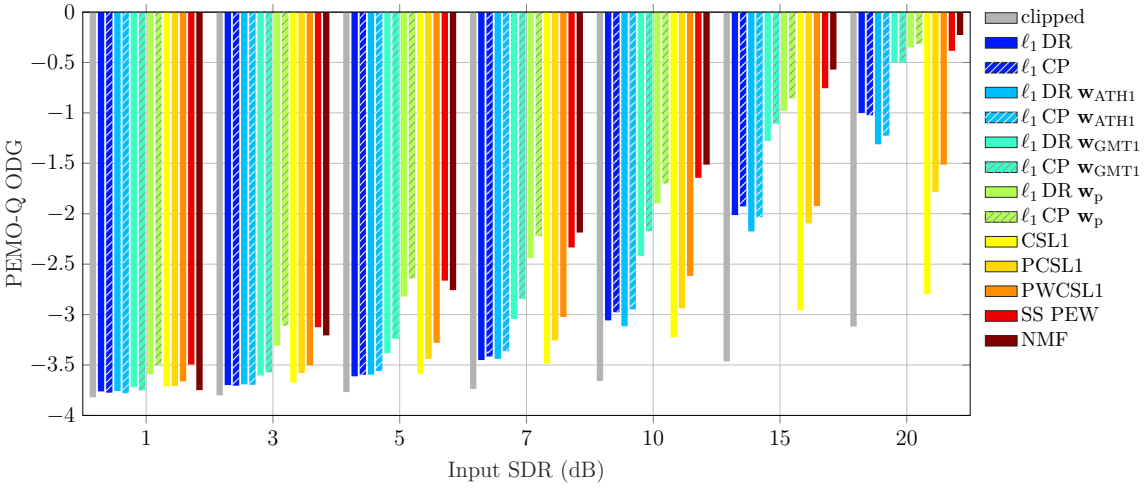


Fig. 5.1: Average declipping performance in terms of PEMO-Q.

To briefly summarize the obtained results, we note that the proposed fully-consistent approaches significantly outperform the CSL1-type algorithms. When weighting is utilized, the analysis variant using the CP algorithm marginally outperforms its synthesis counterpart. The best results obtained by the ℓ_1 CP using the quadratic weights are compatible with the state-of-the-art methods, and in some cases (very low input SDRs) are even slightly better.

6 REPLACING RELIABLE SAMPLES

Some audio declipping methods produce waveforms that do not fully respect the actual process of clipping and allow a deviation from the consistency set Γ (see its definition in (1.2)). In this chapter, the focus is paid to declipping methods producing solutions inconsistent in the reliable part (R-inconsistent), for which it generally holds that $M_{\mathrm{R}}\hat{\mathbf{x}} \neq M_{\mathrm{R}}\mathbf{y}$, where $\hat{\mathbf{x}} \in \mathbb{R}^N$ represents the reconstructed signal obtained by R-inconsistent method, and $\mathbf{y} \in \mathbb{R}^N$ is the clipped observation.

This chapter examines what effect on perception it has if the output of such R-inconsistent methods is pushed towards consistent solutions by postprocessing. First, a simple method based on a straightforward replacement of the reliable samples is described. Consequently, two different solutions are introduced to cope with the negative effects of the basic replacement—one based on audio inpainting and the other exploiting crossfading with the clipped signal.

Basic replacement

The R-inconsistent solutions may be easily turned into consistent by straightforward replacement of the reliable samples from the clipped observation called the basic replacement (BR), formally $M_{\mathrm{R}}\hat{\mathbf{x}} = M_{\mathrm{R}}\mathbf{y}$. However, the main problem of the BR strategy is the risk of creating sharp transitions between the reliable samples (newly replaced by parts of the observed signal) and the rest of the signal (i.e., the reconstructed peaks). Such a nonsmooth phenomenon results in an undesirable occurrence of broadband spectral components, which may have a negative effect on the perceived quality of the restored audio. Nevertheless, the gain in the perceptual quality of the declipped audio obtained by the simple replacement strategy can outweigh the just described disadvantage.

Inpainted replacement

To leverage the knowledge of reliable samples while avoiding the sharp edges at the transitions, a method based on audio inpainting was published in [27]. The main idea of this approach combines the BR method with audio inpainting, such that a number of samples at the beginning and at the end of each clipped section of the signal are "deleted" and then estimated using a selected audio inpainting method, while the "middle" part of the clipped sections along with the (replaced) reliable samples are fixed. This approach is coined Inpainted replacement (IR).

The experiments conducted in the Thesis show that the IR strategy outperforms the basic replacement only for declipping methods that were inferior to prior any replacement. For a priori favorable methods, such as SS PEW, this strategy fails. Apart from the inpainting using plain ℓ_1 minimization approach, the work [27] also introduces a more complex model based on adaptive reliability of the declipped samples. In contrast to the plain IR approach, it allows a nonbinary classification of the reliability of the declipped samples. However, it turns out that this approach only magnifies both the gains and losses obtained by the plain IR.

Crossfaded replacement

Another method designed to suppress the negative effect of sharp transitions caused by the BR method was introduced in [28]. The main principle of this method lies in crossfading the inconsistent declipping solution with the observed signal such that the reconstructed peaks gradually blend into the reliable parts.

Even though the idea of crossfading is fairly simple, there are several options and parameters to choose from. It is the location of the crossfaded region, type of the crossfade, length of the crossfaded section, and a way how to treat segments that are shorter than the predefined length. The Thesis studies the best possible setting of these parameters to achieve the best possible ODG results. It also shows that the CR method provides significantly better results than IR with much lower computational cost and that the CR strategy not only raises the limit of the achievable ODG via SS PEW but also that similar perceptual performance can be reached with significantly fewer iterations.

Results and discussion

The overall PEMO-Q ODG results are illustrated in Fig. 6.1. The results suggest a significant improvement of the reconstruction quality when the crossfaded replacement is applied, especially at medium and high input SDRs. In some cases of very harsh clipping (input SDR of 1 and 3 dB), both replacement strategies can decrease the ODG score of the declipped signal for some of the methods. Nevertheless, the CR technique dominates for high input SDRs and usually provides better results than BR. SS PEW even with applied CR strategy did not outperform the NMF. However, the ODG difference between the two was significantly reduced after the application of CR, with a much lower computational cost.

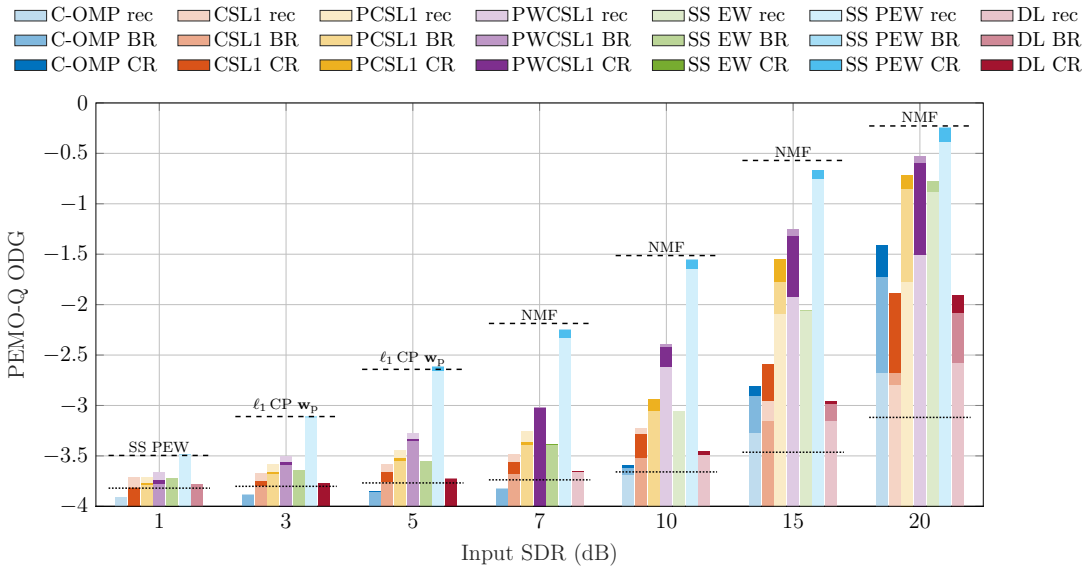


Fig. 6.1: Average PEMO-Q ODG values for inconsistent restoration (lightest color shade), BR strategy (medium shade) and CR strategy (darkest shade). Dotted lines represent the average ODG value of the clipped signals, and the black dashed lines indicate the best ODG result obtained by the methods from the previous chapters.

7 AUDIO DEQUANTIZATION ALGORITHMS

This chapter aims at adopting selected sparsity-based methods previously introduced for audio declipping to the problem of audio dequantization. Since the set of feasible solutions is for both clipping and quantization a box-type set, it is possible to use the same algorithms for audio dequantization as for declipping, and the only difference is in the projection step.

First, the consistent ℓ_1 minimization approach to audio dequantization is utilized and solved via the DR algorithm for the synthesis variant and the CP algorithm for the analysis variant. Allowing some deviation from the feasible set Γ leads to an inconsistent approach, that penalizes the deviation from Γ using the squared distance function. The synthesis variant is in the Thesis solved via FISTA (using the gradient of the distance function) and the DR algorithm (using the proximal operator of the distance function). The analysis variant is solved via the CP algorithm, however, two more alternatives (FISTA and DR algorithm) are introduced, using an approximation via the so-called *approximal operator* [29]. Finally, the consistent heuristic ℓ_0-approximation-based algorithms originally developed for audio declipping as the SPADE algorithms are adapted to audio dequantization and coined as Sparse Audio Dequantizer (SPADQ). The mentioned algorithms were published in [30, 31].

Results and discussion

The average PEMO-Q results are illustrated in Fig. 7.1. The three optimization problems are displayed in different colors (consistent ℓ_1 minimization in blue, inconsistent ℓ_1 minimization in orange, and consistent ℓ_0 approximation in yellow). Moreover, synthesis variants use lighter shades, and analysis variants use darker color shades. Different algorithms are distinguished via hatching (CP and S-SPADQ DP use gray hatching, and FISTA uses black hatching).

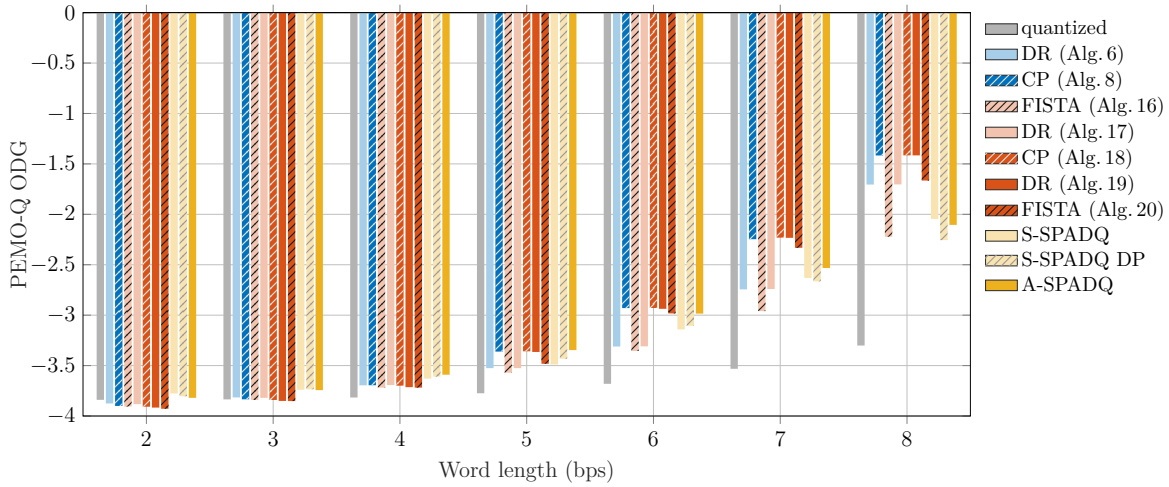


Fig. 7.1: Average dequantization performance in terms of PEMO-Q ODG.

The results show that audio declipping methods can be successfully adapted to dequantization, however, they suggest no clear winner among the tested methods. The SPADQ algorithms perform well for word lengths of 4–7 bps, but they are outperformed by the convex methods for other tested word lengths.

CONCLUSIONS AND PERSPECTIVES

The Doctoral Thesis aimed at the restoration of audio signals corrupted by nonlinear distortions. The focus was primarily devoted to the task of restoring clipped signals, referred to as audio declipping. Nevertheless, a part of the Thesis also dealt with the problem of audio dequantization, which aims at estimating the original signal from its quantized observation.

The first part of the Thesis treated in more detail various sparsity-based audio declipping algorithms. First, we formulated the ℓ_1-relaxed declipping problem in both the synthesis and analysis variants. The synthesis variant was first solved via the Condat–Vũ algorithm, which computed the projections on all three sets R, H, and L separately. Later, the explicit projector onto the whole set of feasible solutions was developed, and thus it was possible to use a simpler and faster Douglas–Rachford algorithm. The analysis variant was solved using the Chambolle–Pock algorithm. Apart from the plain ℓ_1-minimization, we also introduced reweighting of the coefficients to further enhance the sparsity of the solution. It turned out that reweighting significantly improves the results in terms of SDR, especially in the analysis variant. However, the effect was completely reversed in terms of ODG, which suggested that coefficient reweighting is not beneficial for humanocentric audio declipping. Inspired by the promising results of the work by Defraene *et al.* [17], we implemented the Condat–Vũ algorithm to solve the R-inconsistent ℓ_1 minimization-based problem proposed in [17]. Nevertheless, this inconsistency in the reliable part brought no improvement over the consistent variants, and according to psychoacoustically inspired measures, this algorithm lags significantly behind its consistent counterparts. Furthermore, we focused on the ISTA-type declipping algorithm utilizing Social Sparsity. We used the implementation kindly provided by M. Kowalski and slightly accelerated its convergence. The results confirmed those from the original paper [19] that using Persistent Empirical Wiener produces superior restoration quality. The SS PEW algorithm achieved the best results in terms of SDR and performed very well in terms of PEAQ and PEMO-Q. Finally, we examined the SPADE algorithms originally presented in [12]. We reimplemented the algorithms and enhanced their performance by altering the hard thresholding step to respect the conjugate structure of DFT and by utilizing the developed projection lemma, we managed to significantly accelerate the synthesis variant S-SPADE, which, however, turned out not to fully respect the ADMM scheme. Therefore, we developed a new synthesis variant of the algorithm, which significantly outperformed the original S-SPADE. Both the analysis and the new synthesis variants of the algorithm performed well in terms of all evaluation metrics, however, the A-SPADE tended to achieve marginally better results.

Furthermore, we investigated the possibilities of incorporating psychoacoustic information into audio declipping. The a priori information entered the optimization problem in form of weights, which were used to encourage or suppress certain TF coefficients. While weights inspired by the absolute threshold of hearing did not bring an expected improvement of the perceptual quality, the weights obtained from the global masking threshold (specifically a slightly modified MPEG-1 Psychoacoustic Model 1) improved the declipping results up to 0.5 on the PEAQ ODG scale and even slightly more on the PEMO-Q ODG scale. However,

the best overall results by far were obtained by the parabola-based weights, which aim at suppressing the higher harmonics introduced by clipping while the lower frequencies are preserved. Such an option brought significant improvement of the restoration quality in all used evaluation metrics (up to almost 2 on the PEAQ ODG scale and 1.5 on the PEMO-Q scale) with no additional computational cost over the nonweighted variant. The results obtained by the parabola-weighted analysis variant of the ℓ_1-relaxation problem solved via the Chambolle–Pock algorithm were comparable with the top-performing audio declipping methods such as SS PEW and NMF while being ca 6× and 181× faster, respectively.

Next part of the Thesis dealt with the possibilities of improving the results obtained by the declipping methods inconsistent in the reliable part. A basic method where all the samples in reliable positions are replaced with the samples from the clipped observation was introduced and the perceptual effects of such a replacement were studied. Even though most of the inconsistent declipping methods benefited from such basic replacement, at the same time, a major disadvantage consisting in the risk of creating sharp transitions on the borders of the replaced segments was revealed. To leverage the knowledge of the reliable samples while avoiding the sharp edges at the transitions, two other replacement methods were proposed—one based on audio inpainting and the other on crossfading. The latter turned out to be successful in suppressing the sharp transitions and systematically performed better or at least on par with the basic replacement. Apart from the resulting audio quality, it was also shown that applying the crossfaded replacement method during the declipping algorithm can be used to obtain perceptually satisfying results in fewer iterations.

Finally, we tackled the problem of audio dequantization and showed that audio declipping methods can be easily adapted to solve dequantization by altering their projection step. However, despite the close similarity between declipping and dequantization, it does not hold true that methods successful in declipping perform well in dequantization. For instance, the SPADE algorithms for declipping outperformed most of the ℓ_1 minimization-based approaches but the SPADQ algorithms did not fulfill the expectations and turned out to perform mostly on par or even slightly worse than plain ℓ_1 minimization approaches in terms of perceptually motivated measures. The results also pointed out the predominance of analysis variants of the optimization problems, while no significant difference between the consistent methods and methods allowing a deviation from the feasible set was found. An interesting observation was that algorithms exploiting the proximal operator of the differentiable function tend to outperform the gradient-based methods.

To both support the spirit of reproducible research and to stimulate future research in this area, the source codes of the methods described in this Thesis were made publicly available. The MATLAB implementations of the presented audio declipping algorithms including methods aiming at replacing reliable samples are available at the following GitHub repository:

`https://github.com/rajmic/declipping2020_codes`

and audio dequantization

`https://github.com/zawi01/audio_dequantization`.

For audio declipping, a supplementary web page was created. It contains a more detailed comparison of the audio declipping methods, individual results for each audio excerpt and clipping level, and interactive table of the results with the possibility to listen to the declipped excerpts. This web page is available at:

`https://rajmic.github.io/declipping2020`.

To select the most suitable declipping algorithm facing a real-world restoration task, it is necessary to consider several criteria. Some algorithms tend to perform better at low clipping levels, while others perform better at high clipping levels. The choice of an algorithm thus depends on the input data and the possible requirement of the solution consistency. Nevertheless, the methods based on social shrinkage, nonnegative matrix factorization, ℓ_1 minimization with coefficients weighting, and SPADE algorithms yield results that make them preferred choices. Depending on the application, the computational complexity of the algorithms can be a decisive selection criterion. From this point of view, parabola-weighted ℓ_1 CP, and SPADE are attractive. Very good restoration quality with slightly higher computational complexity represents the FISTA exploiting social sparsity, which can be further improved (or accelerated) by applying the crossfaded replacement strategy. If very high computational time is not an issue, then NMF seems to provide the best quality in terms of perceptual metrics.

Following the work presented in this Thesis, we now foresee some ideas and possible directions of further research in the field. A possible way to improve the results is to combine successful strategies of the various algorithms discussed in this Thesis. For instance, the social sparsity regularizer, the parabola-based weights, or the dictionary learning approach could be combined with SPADE or other algorithms. Since the analysis variant of the optimization problem turned out to perform slightly better, the problem solved by Social sparsity algorithm could be reworked into the analysis form. For audio dequantization, other successful declipping algorithms could be applied, for example, the Social sparsity algorithm.

Even though there is still room for improvement, it seems that purely sparsity-based methods are approaching their limits. In other fields of signal processing like computer vision, speech recognition, audio analysis, and many more, it is possible to notice the success of supervised techniques, especially deep learning-based methods. As mentioned in "State of the art chapter" in the Thesis, recent deep learning approaches to speech declipping [32, 33, 34, 35] and audio dequantization [11] have shown promising results, and it seems that future research will follow this trend. A potential direction is also to combine signal modeling and learning from data using the *unrolling*, or *unfolding* approach based on the recent finding that the structure of proximal algorithms can be unrolled into the form of artificial networks [36].

REFERENCES

[1] 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*, 51(11):1012–1031, 2003.

[2] J. Málek. Blind compensation of memoryless nonlinear distortions in sparse signals. In *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5, Sept 2013.

[3] M. J. Harvilla and R. M. Stern. Least squares signal declipping for robust speech recognition. In *15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, pages 2073–2077, Sept. 2014.

[4] Y. Tachioka, T. Narita, and J. Ishii. Speech recognition performance estimation for clipped speech based on objective measures. *Acoustical Science and Technology*, 35(6):324–326, 2014.

[5] P. Wu, X. Zou, M. Sun, J. He, and X. Zhang. The influence of clipping on the performance of a low bit rate parametric speech coder. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5, Oct 2019.

[6] A. H. Poorjam, M. S. Kavalekalam, L. Shi, J. P. Raykov, J. R. Jensen, M. A. Little, and M. G. Christensen. Automatic quality control and enhancement for voice-based remote Parkinson's disease detection. *Speech Communication*, 127:1–16, 2021.

[7] T. Ikoma. Apparatus for avoiding clipping of amplifier, 1984.

[8] M. Bosi and R. Goldberg. *Introduction to Digital Audio Coding and Standards*. Kluwer Academic Publishers, 2003.

[9] C. Brauer, T. Gerkmann, and D. Lorenz. Sparse reconstruction of quantized speech signals. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5940–5944, March 2016.

[10] C. Brauer, Z. Zhao, D. Lorenz, and T. Fingscheidt. Learning to dequantize speech signals by primal-dual networks: an approach for acoustic sensor networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7000–7004, May 2019.

[11] H.-W. Yoon, S.-H. Lee, H.-R. Noh, and S.-W. Lee. Audio dequantization for high fidelity audio generation in flow-based neural vocoder. In *Proc. Interspeech 2020*, pages 3545–3549, Shanghai, China, Oct. 2020.

[12] S. Kitić, N. Bertin, and R. Gribonval. Sparsity and cosparsity for audio declipping: a flexible non-convex approach. In *LVA/ICA 2015 – The 12th International Conference on Latent Variable Analysis and Signal Separation*, pages 243–250, Liberec, Czech Republic, Aug. 2015.

[13] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. PEAQ – The ITU standard for objective measurement of perceived audio quality. *The Journal of the Audio Engineering Society*, 48(1/2):3–29, January/February 2000.

[14] 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 Proc.*, 14(6):1902–1911, November 2006.

[15] P. Záviška, P. Rajmic, A. Ozerov, and L. Rencker. A survey and an extensive evaluation of popular audio declipping methods. *IEEE Journal of Selected Topics in Signal Processing*, 15(1):5–24, 2021.

[16] A. J. Weinstein and M. B. Wakin. Recovering a clipped signal in sparseland. *Sampling Theory in Signal and Image Processing*, 12(1):55–69, 2013.

[17] B. Defraene, N. Mansour, S. D. Hertogh, T. van Waterschoot, M. Diehl, and M. Moonen. Declipping of audio signals using perceptual compressed sensing. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2627–2637, Dec 2013.

[18] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming. `http://cvxr.com/cvx`.

[19] K. Siedenburg, M. Kowalski, and M. Dorfler. Audio declipping with social sparsity. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1577–1581. IEEE, 2014.

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

[21] P. Záviška, P. Rajmic, O. Mokrý, and Z. Průša. A proper version of synthesis-based sparse audio declipper. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 591–595, Brighton, United Kingdom, May 2019.

[22] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley. A constrained matching pursuit approach to audio declipping. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 329–332, 2011.

[23] L. Rencker, F. Bach, W. Wang, and M. D. Plumbley. Consistent dictionary learning for signal declipping. In *Latent Variable Analysis and Signal Separation*, pages 446–455. Springer International Publishing, 2018.

[24] Ç. Bilen, A. Ozerov, and P. Pérez. Audio declipping via nonnegative matrix factorization. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, Oct 2015.

[25] 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*, 34(2):317–330, 4 1986.

[26] P. Záviška, P. Rajmic, and J. Schimmel. Psychoacoustically motivated audio declipping based on weighted l1 minimization. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, pages 338–342, Budapest, Hungary, July 2019.

[27] O. Mokrý and P. Záviška. Inconsistent audio declipping performance enhancement based on audio inpainting. In *Proceedings of the 27th Conference STUDENT EEICT 2021*, pages 596–600. Brno University of Technology, Faculty of Electrical Engineering and Communication, June 2021.

[28] P. Záviška, P. Rajmic, and O. Mokrý. Audio declipping performance enhancement via crossfading. *Signal Processing*, 192:108365, 2022.

[29] O. Mokrý and P. Rajmic. Approximal operator with application to audio inpainting. *Signal Processing*, 179:107807, 2021.

[30] P. Záviška and P. Rajmic. Sparse and cosparse audio dequantization using convex optimization. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, pages 216–220, July 2020.

[31] P. Záviška, P. Rajmic, and O. Mokrý. Audio dequantization using (co)sparse (non)convex methods. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 701–705, Toronto, Canada, 2021.

[32] F. Bie, D. Wang, J. Wang, and T. F. Zheng. Detection and reconstruction of clipped speech for speaker recognition. *Speech Communication*, 72:218 – 231, 2015.

[33] W. Mack and E. A. P. Habets. Declipping speech using deep filtering. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 200–204, Oct 2019.

[34] H. B. Kashani, A. Jodeiri, M. M. Goodarzi, and S. G. Firooz. Image to image translation based on convolutional neural network approach for speech declipping, 2019, 1910.12116.

[35] A. A. Nair and K. Koishida. Cascaded time + time-frequency unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7153–7157, Toronto, Canada, June 2021.

[36] V. Monga, Y. Li, and Y. C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

Curriculum Vitæ

Pavel Záviška

Affiliation

Date of birth:	22. 8. 1993
Address:	Pionýrská 530, Moravský Krumlov, Czech Republic
E-mail:	Pavel.Zaviska@vut.cz, zaviskapavel@gmail.com
Tel:	+420 728 200 507
WWW:	`www.vut.cz/lide/pavel-zaviska-154913` `www.linkedin.com/in/pavel-záviška-010a61188`

QUALIFICATION AND PROFESSIONAL CAREER

Qualification

2017 – 2022 (planned)	Ph.D. in Teleinformatics, Brno University of Technology, Faculty of Electrical Engineering and Communication. Doctoral thesis: *Audio signal declipping and dequantization using sparsity-based methods*
2015 – 2017	MSc. in Telecommunication and Information Technology, Brno University of Technology, Faculty of Electrical Engineering and Communication, Master's thesis: *Audio restoration based on sparse signal representations*
2012 – 2015	BSc. in Teleinformatics, Brno University of Technology, Faculty of Electrical Engineering and Communication, Bachelor's thesis: *Methods for increasing bit-depth in images*

Professional career

2020 – present	Signal processing engineer at the CzechOS company
2017 – present	Junior researcher at the Brno University of Technology

RESEARCH PROJECTS

2020 – 2022	20-29009S Perceptually motivated restoration of highly degraded audio signals
2020 – 2021	FEKT-S-20-6291 Multidisciplinární analýza zvukových a obrazových signálů s použitím moderních technik číslicového zpracování signálů a strojového učení
2017 – 2019	GA17-33798L Restoration of lost information in digital signals
2017 – 2019	FEKT-S-17-4476 Multimodální zpracování nestrukturovaných dat s využitím strojového učení a sofistikovaných metod analýzy signálů a obrazů
2015 – 2016	7AMB15AT033 Moderní metody restaurace digitálních audiosignálů

PROFESSIONAL ACTIVITIES

Scientific internships

October 2019	Institut für Schallforschung (ISF), Vienna, Austria
December 2018	Institut de recherche en informatique et systèmes aléatoires (IRISA), Rennes, France
November 2017	Institut für Schallforschung (ISF), Vienna, Austria

Designated reviewer

International Conference on Telecommunications and Signal Processing
International Journal of Advanced Technology in Engineering and Science
IEEE Journal of Selected Topics in Signal Processing

Teaching activities

2018 – 2020	Introduction to Computer Typography and Graphics
2017 – 2019	Signals and Systems Analysis
2017 – 2018	Digital Signal Processing

OTHER QUALIFICATIONS AND KNOWLEDGE

Language knowledge

Czech language (native speaker)
English language (level C1)
German language (level A2)

Certifications

2019	MATLAB III – performance enhancement, GUI, and OOP; Humusoft
2016	MikroTik MTCNA certificate
2015	Firewall Technologies; AT&T Training Center

Awards

July 2020	Best paper award at TSP 2020 conference
June 2017	Master's degree with honors
June 2017	Dean's award for outstanding Master's thesis
June 2015	Bachelor's degree with honors
June 2015	Dean's award for outstanding Bachelor's thesis

SUMMARY OF PUBLICATION ACTIVITIES

- Scientific journals with impact factor according to Web of Science: 3
- International conferences indexed in Web of Science or Scopus: 7
- Total number of citations according to Web of Science: 36
- Total number of citations according to Scopus: 44
- H-index according to Web of Science: 3
- H-index according to Scopus: 4
- Number of released products: 6

ABSTRACT

Audio signals are susceptible to various types of quality degradation, with clipping being one of the most common and problematic distortions. This Thesis addresses the restoration of audio signals corrupted by nonlinear distortions and presents the contribution in the field of sparsity-based audio restoration algorithms, with the main focus on audio declipping and dequantization. The first part of the Thesis deals with the problem of audio declipping and presents several sparsity-based approaches, containing both the original research and adopted algorithms, which have been reimplemented or modified. The performance of the algorithms is evaluated using the Signal-to-Distortion ratio, as well as perceptually motivated metrics of sound quality. Then, attention is paid on incorporating psychoacoustic information into declipping by weighting the transform coefficients. Three possible constructions of the weights are presented and it is shown that with correctly chosen weights, it is possible to significantly improve the performance of the algorithms, which achieve state-of-the-art restoration quality with low computational complexity. Special focus is also paid on declipping methods that allow a deviation in the reliable part. In that direction, the Thesis studies the perceptual effects of plain replacement of the reliable samples, then identifies its main weaknesses and introduces methods to compensate the discovered negative effects. It is shown that using this technique, it is possible to enhance the performance of such declipping algorithms without a significant increase in computational complexity. Finally, selected declipping algorithms are adopted to the problem of audio dequantization. The Thesis is accompanied by repositories containing implementations of the presented methods.