Short communication

Audio declipping performance enhancement via crossfading

Pavel Záviška, Pavel Rajmic1,*, Ondřej Mokrý

Signal Processing Laboratory, Department of Telecommunications, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technická 12, 612 00 Brno, Czech Republic

A R T I C L E   I N F O

Article history:
Received 6 August 2021
Revised 30 September 2021
Accepted 14 October 2021
Available online 27 October 2021

Keywords:
Clipping
Crossfade
Declipping
Optimization
Sparsity
Reliable samples

A B S T R A C T

Some audio declipping methods produce waveforms that do not fully respect the actual process of clipping and allow a deviation on the reliable samples. This article reports what effect on perception it has if the output of such “inconsistent” methods is pushed towards “consistent” solutions by postprocessing. We first propose a simple sample replacement method, then we identify its main weaknesses and propose an improved variant. The experiments show that the vast majority of inconsistent declipping methods significantly benefit from the proposed approach in terms of objective perceptual metrics. In particular, we show that the SS PEW method based on social sparsity combined with the proposed method performs comparable to top methods from the consistent class, but at a computational cost of one order of magnitude lower.

© 2021 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

Clipping is nonlinear distortion of signals, occurring when there is a lack of available dynamic range. The peak values of a signal are clipped (saturated). In audio, this leads to undesirable, unpleasant perceptual artifacts. Typically, the best option is to avoid clipping beforehand. In cases where clipping can no longer be prevented, there is a need for means of estimating the original signal. Such a process is called declipping.

There are numerous methods for the declipping of signals. Since it is clearly an ill-conditioned task, any declipping method must build on some assumption about the characteristics of a signal. In audio, which is the focus of this article, various methods are based on Bayesian modeling [1], on the autoregressive hypothesis [2], and on low-rank assumptions for matrices [3,4], but most of them are based on the sparsity of signal coefficients with respect to a suitable time-frequency transform [5,6]. For more references, see the declipping survey [7] and a recent overview [8].

For this article, it will be sufficient to distinguish between methods that produce declipping solutions that are consistent in the reliable part of the signal and methods that do not do so. Here, the reliable part corresponds to signal samples that have not been affected by clipping, i.e., a set of samples that fitted within the prescribed dynamic range. The term solution refers to the whole signal in the time domain, even though some of the methods might focus exclusively on reconstructing the clipped samples. For illustration, Fig. 1 shows an observed clipped signal, together with the original and two declipping solutions; one of them is consistent (in the reliable part) and the other is not. We will use the abbreviated reference R-consistent for signals consistent in the reliable part, as well as for methods that produce solutions consistent in the reliable part.

To formalize the concepts, denote \( y \in \mathbb{R}^N \) the observed clipped signal and \( \pm \theta_c \) the clipping thresholds limiting the samples (as in Fig. 1). We can distinguish three sets of samples: \( M_Ry \) are the reliable observed samples, \( M_Hy \) are the samples clipped from above (i.e., equal to \( \theta_c \)), and \( M_Ly \) are the samples clipped from below (i.e., equal to \( -\theta_c \)), where we used the selection masks denoted \( M_R, M_H, \) and \( M_L \). Consistency of the solution \( x_{\text{declip}} \) means that

\[
x_{\text{declip}} \in \Gamma = \{ x \in \mathbb{R}^N \mid M_Rx = M_Ry, M_Hx \geq \theta_c, M_Lx \leq -\theta_c \}.
\] (1)

whereas for R-consistency, the condition \( M_Rx_{\text{declip}} = M_Ry \) is sufficient.

If the signal under treatment is contaminated by noise, insisting on R-consistency goes hand in hand with overfitting. This study, however, assumes noiseless signals, in which case the described declipping inconsistency is in conflict with the observation and with the clipping model. On the other hand, inconsistent methods are typically faster, see for instance Section V.D of [7]. What would happen if, for R-inconsistent methods, the observed reliable sam-

1 Corresponding author.

E-mail addresses: pavel.zaviska@vut.cz (P. Záviška), pavel.rajmic@vut.cz (P. Rajmic), xmokry12@vut.cz (O. Mokrý).

* The work of Pavel Záviška, Pavel Rajmic, and Ondřej Mokrý was supported by the Czech Science Foundation (GAČR) Project No. 20-290095.
samples were taken and put in their place in the current R-inconsistent solution? Such a step would make the solution R-consistent and compatible with the clipping model in the reliable part. Certain signal fidelity measures would improve automatically, such as the signal to distortion ratio (SDR),
\[
\text{SDR}(u, v) = 2 \log_{10} \frac{\|u\|}{\|u - v\|},
\]
when evaluated on the whole original and declipped signals as SDR($x_{\text{orig}}$, $x_{\text{dec}}$). In audio, however, more attention is paid to the perceptual quality, and it is known that the SDR value seldom predicts it well. Thus, the interest of this article lies in exploring how the perceptual quality of the signal reconstruction changes when inconsistent samples in the solution are reconsidered with the help of reliable samples.

This article can be understood as a follow-up to the recent audio declipping survey [7] since the R-inconsistent methods included therein are used as the material for evaluation here.

Section 2 starts with a straightforward replacement of samples, and it shows that such a simple act indeed improves the objective perceptual metrics. On the other hand, the section also reveals the weaknesses of this simple method, leading to the presentation of a more inventive treatment, which still remains conceptually simple and computationally cheap. Section 3 then discusses in greater detail a particular inconsistent algorithm, SS PEW. The section answers the question whether the proposed techniques can improve this already successful algorithm such that it outperforms the state-of-the-art audio declipping algorithms.

2. What shall we do with the reliable samples?

In this section, we will examine to what extent the signal reconstruction quality is affected when (re)using the reliable samples. Note again that such a postprocessing step makes sense only for declipping methods that produce solutions inconsistent in the reliable part (R-inconsistent). Specifically, we will follow up on the survey [7] and consider R-inconsistent methods included therein: The Constrained Orthogonal Matching Pursuit (C-OMP [9]), Plain and Perceptually-motivated Compressed Sensing L1 (CSL1, PCSL1 [10]), Parabola-weighted Compressed Sensing L1 (PWCSL1 [7]), Declipping with Empirical Wiener shrinkage and Social Sparsity Declipping with Persistent Empirical Wiener (EW, PEW [11]), and Dictionary Learning (DL [12]). The detailed setting of the declipping methods is identical to the setting used in the declipping survey [7].

Also, the audio dataset used in our experiments is identical to the one used in the survey [7]. The corpus consists of 10 musical excerpts in mono, sampled at 44.1 kHz with a bit depth of 16 bps and with an approximate duration of 7 seconds. The samples originally come from the EBU SQAM dataset [1] and they cover a range of audio characteristics. Per each signal, 7 degraded versions were obtained by (artificial) clipping using 7 different pairs of the symmetric thresholds $-\theta$, $\theta$. These were computed according to prescribed input signal-to-distortion ratios, defined as SDR($x_{\text{orig}}$, $x_{\text{clip}}$), and ranging from 1 dB to 20 dB.

For the objective perceptual quality prediction, the PEAQ (Perceptual Evaluation of Audio Quality) was used. The output of PEAQ is the Objective Difference Grade (ODG), which predicts the human rating of the difference between the degraded (or reconstructed) and the reference signals. Possible values from $-4$ to $0$ correspond to the scale “very annoying”—“annoying”—“slightly annoying”—“perceptible, but not annoying”—“imperceptible”. We use the MATLAB code [3] implemented according to the revised version of PEAQ (BS.1387-1) [13].

As another evaluation metric taking the human auditory system into account, the PEMO-Q [4] for Matlab was used. PEMO-Q computes the perceptual similarity measure (PSM), which can be mapped onto the ODG score as used in the PEAQ.

2.1. Basic replacement strategy (RR for replace reliable)

The basic approach has already been indicated: the samples of the reconstructed signal are simply substituted with the reliable samples in all parts where the signal has not been clipped. Fig. 2 illustrates such an approach. At the same time, this figure reveals the main, not unexpected problem of RR, which is the risk of creating sharp transitions between the newly recreated reliable parts and the rest of the signal. Perceptually, a non-smooth phenomenon like this results in an undesirable occurrence of broadband spectral components.

Nevertheless, the gain in the perceptual quality of the declipped audio obtained by the simple replacement strategy can outweigh the just described disadvantage, as visible in Fig. 3. This figure shows the average PEAQ ODG improvement obtained using the basic replacement strategy, i.e., the difference of the new ODG value and the original ODG value. The average is computed over the ODG values of individual excerpts. The depicted values of $\Delta$ PEAQ ODG suggest that the RR strategy is, for some declipping methods, not beneficial at low input SDR levels (i.e., a low number of reliable samples). Generally, the improvement grows with increasing input SDR, even up to two grades of ODG in the case of the PCSL1 algorithm at 20 dB input SDR. Note that the replacement (and thus the
improvement) comes essentially for free from the computational viewpoint.

2.2. Advanced strategy (CR for crossfade reliable)

Clearly, a smarter strategy of enhancing the perceived quality should avoid sharp transitions on the borders between the reliable and the clipped regions. Such a method can stem from the simple one, and the desired effect is achieved by crossfading the inconsistent declipping solution with the observed signal such that the reconstructed signal gradually blends into the reliable parts of the signal. See [15] for an example of similar use of crossfading in a signal reconstruction problem.

The crossfading transition can be performed either in the clipped part, in the reliable part, or in the middle affecting both parts. From the three options, only the transition in the reliable part is not affected by the initially clipped samples, which suggests the most psychoacoustically pleasing result. However, some of the reliable samples are altered this way and the resulting signal is thus not guaranteed to be R-consistent anymore.

There are several types of crossfades used in practice [16]. In this paper, we examined the simple linear crossfades, which is usually used for highly correlated material, and the smooth crossfades, whose curve is governed by the squared sine.

An important parameter of a crossfade is the length of the crossfaded section, which determines the number of modified samples. In the case of transition in the reliable part, the longer the transition is, the smoother one signal blends into the other; however, more samples will then differ from the ground truth. Hand in hand with specifying the crossfade length, it must be decided how to treat segments that are shorter than the predefined length. These can be either ignored (keeping the samples from the restored signal unaltered), replaced using the RR strategy or the length of the crossfade can be adaptively shortened to fit the length of the processed segment. The adaptive shortening approach simply computes the largest possible length of the crossfade for each transition, such that the crossfaded sections for adjacent transitions do not overlap.

Different setups for the CR strategy are thus available. Experimenting with the possible combinations showed convincingly that the transition in the reliable part produced the best perceptual results according to both PEAQ and PEMO-Q, as expected. But in terms of the width and shape of the crossfades and in terms of the way of treating the short segments, the results vary according to the evaluation metric. PEAQ seems to respond positively to the smooth crossfade and to ignoring the processing of shorter segments, while PEMO-Q favors the linear transition and adaptive shortening of the short transitions. Nonetheless, the differences between these setups are negligible (up to 0.1 on the ODG scale).

To allow a concluding statement about the impact of the proposed CR strategy, the same experiment as with the RR strategy was conducted. As for the setup, the transition in the reliable part and an 8-sample-long smooth crossfade with adaptive shortening were used, based on the above discussion. Fig. 4 displays the average PEAQ ODG and PEMO-Q ODG values. The individual declipping algorithms are distinguished using different bar colors. Within a single bar, the lightest shade represents the quality of the originally clipped, inconsistent signal. The respective medium shade marks the results of the RR strategy, and finally, the darkest shade corresponds to CR. Note that the graph has to be read in a cumulative manner, i.e., the particular ODG grade is expressed by the top level of each shaded section of the corresponding bar. Additionally, the black dotted lines represent the average ODG value of the clipped signals, and the black dashed lines indicate the best ODG result obtained in the survey [7], which were mostly results of the non-negative matrix factorization (NMF) and the analysis variant of parabola-weighted $\ell_1$-minimization (PW$\ell_1$ CR).

The PEAQ results in Fig. 4 (top) suggest a significant improvement of the reconstruction quality when the advanced crossfading method is applied, especially at medium and high input SDRs. The CR method always performs better or at least on par with the RR strategy. However, in some cases of very harsh clipping (input SDR of 1 and 3 dB), both replacement methods can decrease the PEAQ score of the declipped signal. The results displayed in Fig. 4 (bottom) indicate PEMO-Q being more conservative in comparison to PEAQ, but even in this case the CR strategy usually leads to an improvement, compared to RR.

3. A closer look at SS PEW

In the audio declipping survey [7], the method based on the so-called social sparsity (SS) with persistent empirical Wiener PEW shrinkage [11,17] ranked high. SS PEW was in fact the overall best-performing method in terms of the SDR. In terms of PEMO-Q, SS PEW was also one of the top-performing methods, being outperformed only slightly by two of the competitors, namely the parabola-weighted $\ell_1$ minimization (PW$\ell_1$, [6]) and the non-negative matrix factorization (NMF, [3]). However, PEAQ placed SS PEW a bit deeper in the racing list. The reader’s subjective judgement can be easily obtained by listening to the audio examples at the survey’s accompanying webpage (see the link in Section 4).

PW$\ell_1$ and NMF are actually fully consistent approaches, and the NMF is computationally more expensive by one order compared to other declipping methods [7, Section V.D]. Since SS PEW produces R-inconsistent solutions, the question arises as to how much we can enhance the declipping quality of SS PEW by using the RR or CR methods. How will it compare with PW$\ell_1$ and NMF then? Besides this, would it even be possible to reduce the computational time by computing fewer SS PEW iterations, while still being able to reach the competitive quality of reconstruction with the help of the RR or CR strategies?

The algorithms based on social sparsity approximate the solution to the following optimization problem:

$$\min_{z, \theta} \left\{ \frac{1}{2} \| M_R Dz - M_R y \|_2^2 + \frac{1}{2} \| h(M_R Dz - M_R \theta z) \|_2^2 + \frac{1}{2} \| h(-M_L Dz - M_L \theta) \|_2^2 + \lambda R(z) \right\}.$$  

(3)

The first term penalizes the $\ell_2$ error in the reliable region, which is captured by the mask operator $M_R$ (cf. the definition of the set $\Gamma$ (1)). The vector $z \in \mathbb{C}^N$ represents coefficients approximating the signal $y \in \mathbb{R}^N$ with respect to a selected time-frequency (TF) transform, here a synthesis operator $D : \mathbb{C}^N \rightarrow \mathbb{R}^N$. The deviation of the solution from the set $\Gamma$ in the clipped regions is penalized using
the hinge function $h$, defined as the identity for its negative arguments, and zero otherwise. Finally, $R$ is a sparsity-enforcing regularizer. The authors of [11] suggest four types of shrinkage operator related to $R$ for use in practical algorithms, and the best-performing in audio declipping turned out to be the Persistent Empirical Wiener (PEW) [7,8,11]:

$$S_{\text{PEW}}(z_{ft}) = z_{ft} \cdot \max \left( 1 - \frac{\lambda}{\|N(z_{ft})\|_2}, 0 \right).$$  \hspace{1cm} (4)

The indexes $t$ and $f$ specify the location of a coefficient in time and frequency, respectively, and $N(z_{ft})$ denotes the TF neighborhood of the coefficient at location $(f,t)$.

The parameter $\lambda$ in (3) balances sparsity and data fidelity. Larger values of $\lambda$ lead to a solution of higher sparsity (meaning fewer coefficients) but also of greater deviance from the solution consistency, and vice versa. To accelerate the overall convergence, the algorithm proposed in [11] implements the adaptive restart strategy [18]; the optimization actually starts with a larger $\lambda$ and it is decreased after every few hundred iterations until the target value of $\lambda$ is reached. This way, outer and inner iterations are recognized.

Figs. 5 and 6 demonstrate the evolution of the SDR computed in the clipped and reliable parts, respectively, across iterations. The plots show that while the SDR continues to grow in the reliable part with the increasing number of iterations, the SDR in the clipped part stabilizes after reaching a certain value. The pictures show the SDR for the particular case when $\lambda = 10^{-4}$, which is in agreement with the setup from [7]; nevertheless, such behavior has also been verified for $\lambda = 10^{-5}$. In other words, from a certain point on, iterations minimize only the difference in the reliable part. This observation actually supports the above-proposed idea of terminating the iterations of SS PEW earlier and applying the RR/CR postprocessing.

---

**Fig. 4.** Average PEAQ (top) and PEMO-Q (bottom) ODG values for inconsistent restoration (lightest color shade), RR strategy (medium shade) and CR strategy (darkest shade). Each group of bars is crossed by a horizontal dotted line; these mark the ODGs of the clipped signals. The dashed lines are present to indicate the best possible ODG results achieved by a method from [7].

**Fig. 5.** Average SDR for SS PEW in the clipped part in the course of iterations, with $\lambda = 10^{-4}$. 

---

P. Záviška, P. Bajnici and O. Mokrý  
Signal Processing 192 (2022) 108365
verify the predictions of PEMO-Q and PEAQ metrics, since these measures were not originally intended to evaluate the quality of clipped and reconstructed audio files.

**Declaration of Competing Interest**

The authors of the article being submitted all declare that they have seen and approved the manuscript and have no conflicts of interest to disclose.

**Credit authorship contribution statement**

**Pavel Záviška:** Investigation, Software, Visualization, Writing – original draft, Writing – review & editing. **Pavel Rajmic:** Methodology, Supervision, Writing – original draft, Writing – review & editing. **Ondrej Mokrý:** Investigation, Writing – original draft, Writing – review & editing.

**References**

[1] W. Fong, S. Godsell, Monte carlo smoothing for non-linearly distorted signals, in: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 6, 2001, pp. 3997–4000.

[2] A. Dahimene, M. Noureddine, A. Azar, A simple algorithm for the restoration of clipped speech signal, Informatica 32 (2008) 183–188.

[3] Ç. Bilen, A. Özerov, P. Pérez, Audio declipping via nonnegative matrix factorization, in: 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1–5.

[4] T. Takahashi, K. Uruma, K. Konishi, T. Furukawa, Block adaptive algorithm for signal declipping based on null space alternating optimization, IEICE Trans Inf Syst (1) (2015) 206–209, doi:10.1587/transinf.2014EDL81098.E98.D

[5] S. Kretí, N. Bertin, R. Gribonval, Sparsity-based audio declipping: a flexible non-convex approach, in: LVA/ICA 2015 – The 12th International Conference on Latent Variable Analysis and Signal Separation, 2015, pp. 243–250.

[6] P. Záviška, P. Rajmic, J. Schimmel, Psychoacoustically motivated audio declipping based on weighted L1 minimization, in: 2019 42nd International Conference on Telegraph Communications and Signal Processing (TSP), 2019, pp. 334–342.

[7] P. Záviška, P. Rajmic, A. Özerov, L. Rencker, A survey and an extensive evaluation of popular audio declipping methods, IEEE J Sel Top Signal Process 15 (1) (2021) 5–24, doi:10.1109/JSTSP.2020.3042071.

[8] C. Galtier, S. Kretí, R. Gribonval, N. Bertin, Sparsity-based audio declipping methods: selected overview, new algorithms, and large-scale evaluation, IEEE/ACM Trans Audio Speech Lang Process 29 (2021) 1174–1187, doi:10.1109/TASLP.2021.3059264.

[9] A. Adler, V. Eniya, M. Jafari, M. Elad, R. Gribonval, M. Plumbley, A Constrained Matching Pursuit Approach to Audio Declipping, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 329–332, doi:10.1109/ICASSP.2011.5946407.

[10] B. Debraene, N. Mansour, S.D. Hertogh, M. Diehl, M. Moonen, T.V. Waterschoot, Declipping of audio signals using perceptual compressed sensing, IEEE Trans Audio Speech Lang Process 21 (12) (2013) 2627–2637, doi:10.1109/TASLP.2013.2281570.

[11] K. Siedenburg, M. Kowalski, M. Dörfler, Audio Declipping with Social Sparsity, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1577–1581, doi:10.1109/ICASSP.2014.6853863.

[12] L. Rencker, F. Bach, W. Wang, M.D. Plumbley, Consistent Dictionary Learning for Signal Declipping, in: LVA/ICA 2018 – The 12th International Conference on Latent Variable Analysis and Signal Separation, 2018, pp. 446–455, doi:10.1007/978-3-319-93764-5_4.

[13] P. Kabal, An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality, MMSP Lab Technical Report, Dept. Electrical & Computer Engineering, 2002, McGill University, Tech. Rep.

[14] R. Huber, B. Kollmeier, PEMO-Q—a new method for objective audio quality assessment using a model of auditory perception, IEEE Trans. Audio Speech Language Process. 14 (6) (2006) 1902–1911, doi:10.1109/TASL.2006.883259.

[15] P.A.A. Esquéf, K. Roth, I. Kauppinen, V. Valimäki, Interpolation of Long Gaps in Audio Signals Using the Warped Bung’s Method, in: Proceedings of the 6th Int. Conference on Digital Audio Effects (DAFx-03), 2003.

[16] L. Lupsa-Tataru, Audio cross-fade implementation via rational functions, Technologies Romanian Journal of Applied Sciences and Technology 3 (5) (2021) 39–48, doi:10.47577/technoxim.35.3881.

[17] K. Siedenburg, Persistent Empirical Wiener Estimation with Adaptive Threshold Selection for Audio Denoising, in: Proceedings of the 9th Sound and Music Computing Conference, 2012, pp. 426–433.

[18] B. O’Donoghue, E. Candès, Adaptive restart for accelerated gradient schemes, Foundations of Computational Mathematics 15 (3) (2013) 715–732, doi:10.1007/s10208-013-9150-3.