INCONSISTENT AUDIO DECLIPPING PERFORMANCE ENHANCEMENT BASED ON AUDIO INPAINTING

Ondřej Mokrý and Pavel Záviška
Doctoral Degree Programme (2 and 4), FEEC BUT
E-mail: 170583@vutbr.cz, xzavis01@vutbr.cz

Supervised by: Pavel Rajmic
E-mail: rajmic@vutbr.cz

Abstract: Some of the state-of-the-art audio declipping methods are not consistent with the observed signal, meaning that they do not keep the a priori undegraded (reliable) samples. These samples can be directly substituted in the reconstructed signals, but this may create audible artifacts due to the sharp transitions between the declipped and the substituted parts. We propose two methods based on audio inpainting, which deal with these transitions. As a result, we observe a significant improvement of the PEAQ ODG values, but only for some of the declipping algorithms considered.

Keywords: audio, clipping, declipping, inpainting, sparsity, perceptual quality

1 INTRODUCTION

Clipping is a nonlinear distortion that cuts off signal peaks exceeding the allowed dynamic range $[-\theta_c, \theta_c]$, where the value $\theta_c > 0$ is referred to as the clipping threshold. Formally, for the n-th sample of the signal $\mathbf{x} \in \mathbb{C}^L$, clipping produces the sample $\theta_c \cdot \mathrm{sgn}(x_n)$ for $|x_n| \geq \theta_c$, and keeps the sample x_n otherwise.

The nonsmooth transitions caused by clipping produce higher harmonics that were not originally present in the signal, which is perceived as an unpleasant distortion. Not only clipping degrades the perceptual quality of audio signals [1], it also worsens further processing conditions, such as automatic speech recognition [2] or voice-based Parkinson's disease detection [3].

The reconstruction of a signal from its clipped observation is generally referred to as *declipping*. For a large overview of the declipping methods, see the audio declipping survey [4].

In declipping, it feels natural to require that the declipped samples should exceed the high (θ_c) and low ($-\theta_c$) clipping thresholds and that the samples that were not altered by clipping (i.e., the *reliable* samples) should remain the same. Signals satisfying these conditions form the *consistent set* and methods that find a solution belonging to this set are naturally referred to as *consistent methods*. On the other hand, *inconsistent methods* only penalize the distance of the solution from the consistent set. See Fig. 1 for the comparison of both approaches. The inconsistent methods include Constrained Orthogonal Matching Pursuit (C-OMP [5]), Plain and Perceptually-motivated Compressed Sensing L1 (CSL1, PCSL1 [6]), Parabola-weighted Compressed Sensing L1 (PWCSL1 [4]), Declipping with Empirical Wiener Shrinkage and Social Sparsity Declipping with Persistent Empirical Wiener (SS EW, SS PEW [7]), or Dictionary Learning (DL [8]). The advantages are usually their low computational complexity, and the possibility to fully exploit any prior about the signal structure, such as the sparsity of its time-frequency representation.

However, sticking to the inconsistent solution implies that the knowledge of the reliable samples from the clipped signal is undervalued. This paper, therefore, addresses the issue of enhancing the quality of inconsistent audio declipping methods by replacing the reliable samples.

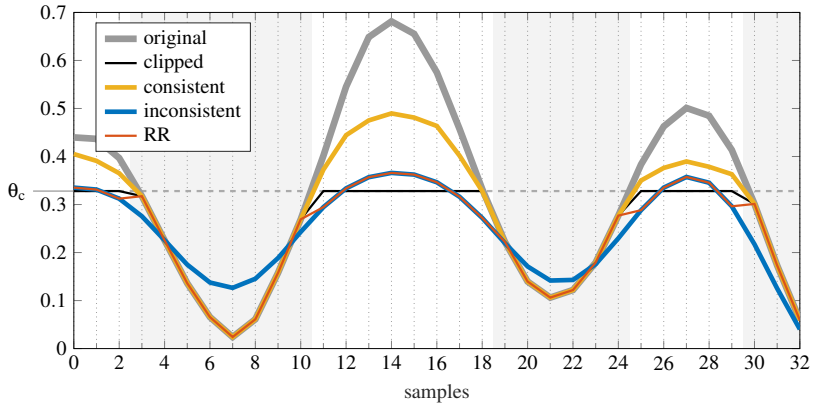


Figure 1: Demonstration of consistent and inconsistent declipping solutions on a short piece of audio signal including the application of the straightforward replacement of the reliable samples (RR).

2 PROBLEM FORMULATION

The easiest way to fully exploit the information stored in the reliable samples is simply to substitute them in place of the inconsistent samples. This procedure naturally increases the signal-to-distortion ratio (SDR) and in a large number of cases also improves the perceptual quality according to the perceptual measures, such as PEAQ [9].

On the other hand, the main problem of this technique is the possibility of creating sharp transitions between the clipped and the replaced reliable parts, as illustrated in Fig. 1. Such a nonsmooth phenomenon creates higher harmonic spectral components and thus degrades the perceived audio quality.

To leverage the knowledge of the reliable samples while avoiding the sharp edges on the transitions, we present a novel method based on audio inpainting. The main idea lies in "deleting" a number of samples in the beginning and at the end of each clipped part and then estimating the values of these samples using an arbitrary inpainting method, while the rest of the clipped parts, as well as the (replaced) reliable samples are fixed. The set of signals meeting such conditions will be denoted Γ. Suitable audio inpainting methods, which make use of the just defined set Γ, are described in the following section.

3 INPAINTING THE TRANSITIONS

Audio signals are typically sparse in a time-frequency (TF) representation, which is caused by their harmonic and short-time stationary nature. Since the problem of finding a signal in Γ is ill-posed (there are infinitely many solutions meeting the conditions), we may regularize it such that the solution is desired to have a sparse TF representation or a TF representation with minimal ℓ_1 norm [10]. The sparsity-based audio inpainting problem can be formulated as one of the following optimization problems:

$$\arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 + g(A^*\mathbf{c}), \tag{1}$$

$$\arg\min_{\mathbf{x}} \|A\mathbf{x}\|_1 + g(\mathbf{x}). \tag{2}$$

The formulations (1) and (2) are referred to as the synthesis and analysis formulations, respectively. The variable $\mathbf{c} \in \mathbb{C}^N$ denotes a TF representation of a signal $\mathbf{x} \in \mathbb{C}^L$. We use a redundant transform, such that the analysis operator $A\colon \mathbb{C}^L \to \mathbb{C}^N$, $N > L$, produces TF coefficients of a signal, and the synthesis operator $A^*\colon \mathbb{C}^N \to \mathbb{C}^L$ reconstructs the signal out of the coefficients. In such a redundant case, the two formulations are not equivalent. Furthermore, we suppose $A^*A = \mathrm{Id}$.

In both formulations, the norm $\|\cdot\|_1$ promotes sparsity of the TF coefficients, and g enforces the signal consistency in the time domain. As indicated above, we may put $g = \iota_\Gamma$, i.e., the indicator function, which takes on zero value for elements belonging to Γ and infinity otherwise.

The problem (1) can be solved via the Douglas–Rachford algorithm [11, Sec. IV]. Keeping the notation of [11], let $f_1 = \|\cdot\|_1$ and $f_2 = g \circ A^*$. Then, the proximal mapping $\operatorname{prox}_{\gamma f_1} = \operatorname{prox}_{\gamma\|\cdot\|_1}$ is the soft thresholding with threshold γ and $\operatorname{prox}_{\gamma f_2} = \mathrm{Id} + A \circ (\operatorname{prox}_{\gamma g} - \mathrm{Id}) \circ A^*$ [11, Tab. I, ix]. When $g = \iota_\Gamma$, $\operatorname{prox}_{\tau g} = \operatorname{prox}_{\iota_\Gamma}$ is the projection onto Γ, i.e., it replaces all the fixed samples and keeps the rest.

The problem (2) can be solved via the Chambolle–Pock algorithm [12, Alg. 1]. Keeping the notation, we put $F = \|\cdot\|_1$ and $G = g$; the linear operator A is taken care of by the algorithm. The procedure makes use of the proximal mappings $\operatorname{prox}_{\sigma F^*}$ and $\operatorname{prox}_{\tau G}$, where the asterisk in F^* denotes the convex conjugate of the function F. It holds $\operatorname{prox}_{\sigma F^*} = \mathrm{Id} - \sigma \operatorname{prox}_{F/\sigma}(\cdot/\sigma)$ [12, Eq. (6)]. In our settings, $\operatorname{prox}_{\sigma\|\cdot\|_1^*} = \mathrm{Id} - \sigma \operatorname{prox}_{\|\cdot\|_1/\sigma}(\cdot/\sigma)$, i.e., the soft thresholding operator is used with the threshold $1/\sigma$.

4 ADAPTIVE RELIABILITY OF THE DECLIPPED SAMPLES

The model in the previous section can be understood such that we trust the originally reliable samples and also some of the declipped samples, while we do not trust the declipped samples near the transitions *at all*. A natural generalization would be that the reliability of the declipped samples is not only binary. This can be implemented in the function g. If we take the declipped signal and simply substitute the reliable samples, we obtain the signal $\hat{\mathbf{x}}$. Per each declipped sample n, we may put

$$g_n(x_n) = \begin{cases} w_n \cdot |x_n - \hat{x}_n|^2/2 & \text{for } \hat{x}_n \text{ declipped,} \\ \iota_{\{\hat{x}_n\}}(x_n) & \text{for } \hat{x}_n \text{ reliable,} \end{cases} \qquad g(\mathbf{x}) = \sum_{n=1}^{L} g_n(x_n), \tag{3}$$

where $w_n \geq 0$ are the *reliability weights*: the greater the w_n, the more we trust the sample $\hat{x}_n$. To apply the previously described proximal algorithms, we use the proximal mapping of the function g, which is defined elementwise as [13, Thm. 6.6, Ex. 6.65]:

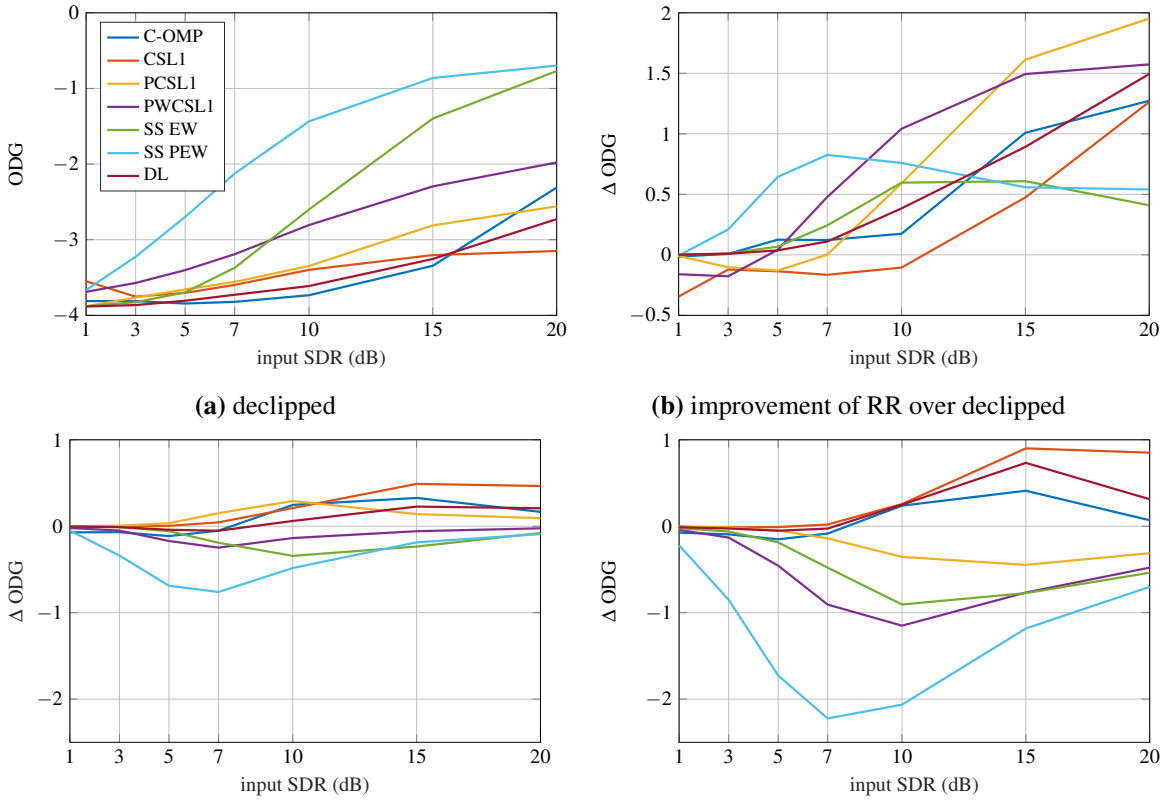
$$x_n \mapsto \frac{w_n}{w_n + 1}\hat{x}_n + \frac{1}{w_n + 1}x_n. \tag{4}$$

The simple inpainting approach from Sec. 3 fits this model with $w_n = \infty$ for the fixed samples and $w_n = 0$ for the transitions to be filled. For the generalized method, we suggest using a bump function in each declipped segment to assign the weights w_n, since the transitions are less trusted than the declipped samples in the middle of each segment.

5 EXPERIMENTS AND RESULTS

For the experiments, 10 audio excerpts sampled at 44.1 kHz with their approximate length of 7 seconds were used. These signals were artificially clipped as defined in the Introduction with 7 different threshold based on the input SDR. The restored signals were obtained using the inconsistent methods mentioned in the Introduction and included in the survey [4] (C-OMP, CSL1, PCSL1, PWCSL1, SS EW, SS PEW, and DL). Except for C-OMP and DL, the methods are based on the convex sparsity-inspired formulation (1) with g measuring the distance from the consistent set of the declipping problem. C-OMP aims at approximating the non-convex sparse synthesis problem in a greedy way, constraining the solution to lie near the consistent set. Finally, DL enhances the sparsity based approach by searching not only for the signal but also the best operator A^* to allow for sparser solution.

As the TF transform A, we used the discrete Gabor transform, implemented for MATLAB in the toolbox LTFAT [14]. Two settings of the transform were tested: The first system was defined by the Hann window of length 8192 samples, window shift 2048 and 8192 frequency channels. The



(a) declipped

(b) improvement of RR over declipped

(c) improvement of INP1 over RR

(d) improvement of INP2 over RR

Figure 2: Average PEAQ performance of the inconsistent methods and the replacing strategies.

second system used the same window shape, but the numerical parameters were 4096, 1024 and 4096, respectively. Both systems were scaled to satisfy the relation $A^*A = \mathrm{Id}$.

In the experiment, the Douglas–Rachford algorithm was applied with $\gamma = 1$ and $\lambda_n = 1$ for all n. It stopped either after 500 iterations or when the relative change of the solution dropped below $5 \cdot 10^{-4}$. Similarly, the Chambolle–Pock algorithm was applied with $\tau = \sigma = 1$ and the same stopping criteria.

The perceptual quality of audio signals was evaluated using PEAQ (Perceptual Evaluation of Audio Quality) [9], specifically the free MATLAB implementation [15]. PEAQ predicts the degradation using a ODG scale ranging from -4 (very annoying) to 0 (imperceptible).

Fig. 2 presents the PEAQ ODG values of the declipped signal, prior to any replacement, and the PEAQ ODG improvement (Δ ODG) for different postprocessing strategies: basic replacement (RR), the simple inpainting method from Section 3 (INP1) and the modified inpainting method from Section 4 (INP2). Only a selection of all the possible settings is plotted, namely INP1 used the synthesis formulation and the TF transform with longer window, INP2 used the analysis formulation and the TF transform with shorter window. Per each declipped segment of length D, INP1 filled d samples at each of the two transitions, with $d = \max(\lfloor D/3 \rfloor, 10)$. INP2 was applied with the weights according to the function $10\sin(t)^{1/2}$, with t representing D evenly spaced samples from 0 to π.

The most significant observation here is that the inpainting-based strategies outperform the basic replacement strategy in case of the methods that were inferior prior to any replacement. On the other hand, these strategies fail to enhance the a priori favorable methods, such as SS PEW.

Furthermore, comparing the bottom graphs leads to a clear conclusion that INP2 magnifies both the gains and the losses of INP1.

6 CONCLUSION

We have presented two inpainting-based methods, which serve as a postprocessing procedure for inconsistent audio declipping. The goal was to maximally exploit the reliable samples of the original declipping problem while avoiding sharp transitions in the signal after a simple substitution. The proposed methods enhance the quality of the resulting audio signal based on PEAQ ODG, but only for some of the declipping methods involved. Although the improvement seems to be rather significant in these cases (up to 1 degree on the ODG scale), the inpainting methods do not improve the reconstruction quality, compared to the basic replacement strategy, for the rest of the methods.

ACKNOWLEDGEMENT

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

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," *Journal of the Audio Engineering Society*, 2003.

[2] Y. Tachioka, T. Narita, and J. Ishii, "Speech recognition performance estimation for clipped speech based on objective measures," *Acoustical Science and Technology*, 2014.

[3] A. H. Poorjam, et al., "Automatic quality control and enhancement for voice-based remote Parkinson's disease detection," *Speech Communication*, 2021.

[4] 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*, 2021.

[5] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley, "A constrained matching pursuit approach to audio declipping," in *IEEE ICASSP*, 2011.

[6] B. Defraene, et al., "Declipping of audio signals using perceptual compressed sensing," *IEEE Trans. Audio, Speech, and Language Processing*, 2013.

[7] K. Siedenburg, et al., "Audio declipping with social sparsity," in *IEEE ICASSP*, 2014.

[8] L. Rencker, F. Bach, W. Wang, and M. D. Plumbley, "Consistent dictionary learning for signal declipping," in *Latent Variable Analysis and Signal Separation*. Springer, 2018.

[9] T. Thiede, et al., "PEAQ – The ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, 2000.

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

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

[12] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, 2011.

[13] A. Beck, *First-Order Methods in Optimization*. SIAM, 2017.

[14] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The Large Time-Frequency Analysis Toolbox 2.0," in *Sound, Music, and Motion*. Springer, 2014.

[15] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," MMSP Lab Technical Report, McGill University, 2002.