ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095928

UPGLADE: UNPLUGGED PLUG-AND-PLAY AUDIO DECLIPPER BASED ON CONSENSUS EQUILIBRIUM OF DNN AND SPARSE OPTIMIZATION

Tomoro Tanaka[†], *Kohei Yatabe*[‡], *Yasuhiro Oikawa*[†]

[†] Waseda University, Tokyo, Japan [‡] Tokyo University of Agriculture and Technology, Tokyo, Japan

ABSTRACT

In this paper, we propose a novel audio declipping method that fuses sparse-optimization-based and deep neural network (DNN)–based methods. The two methods have contrasting characteristics, depending on clipping level. Sparse-optimization-based audio declipping can preserve reliable samples, being suitable for precise restoration of small clipping. Besides, DNN-based methods are potent for recovering large clipping thanks to their data-driven approaches. Therefore, if these two methods are properly combined, audio declipping effective for a wide range of clipping levels can be realized. In the proposed method, we use a framework called consensus equilibrium to fuse the above two methods. Our experiments confirmed that the proposed method was superior to both conventional sparse-optimization-based and DNN-based methods.

Index Terms— Audio declipping, deep learning, fixed-point algorithm, consensus optimization, hard clipping.

1. INTRODUCTION

Clipping is one of the most common audio signal distortions that may occur in audio acquisition. This paper focuses on *hard clipping*, where an observed signal gets truncated at a clipping level $\tau > 0$, as shown in Fig. 1. Clipping degrades audio quality [1, 2] and has negative effects on subsequent processing [3,4]. Therefore, recovery of clipped signals, or *audio declipping*, is desired.

While audio declipping has been tackled by several methods thus far, this paper focuses on the mainstream methods: sparse-optimization-based methods [5–10] and deep neural network (DNN)–based methods [11–13]. They have almost contrasting characteristics, which are summarized in Fig. 2. First, let us discuss sparse-optimization-based methods. They can preserve reliable samples (which are in gray backgrounds in Fig. 1) via a time-domain constraint in the optimization problem and can directly use their information. Thus, sparse-optimization-based methods effectively recover signals with small clipping. However, since they assume a general property of audio signals, sparsity in the time-frequency (T-F) domain, sparse-optimization-based methods indirectly tackle audio declipping. Therefore, they might not work well when clipping is severe.

Next, let us move on to DNN-based methods, which have been emerging lately. DNN extracts relevant information from datasets and then utilizes it for inference. In audio declipping, a DNN would learn the information about clipping (e.g., when an audio signal is clipped, how it would turn to be) and use it to directly tackle audio declipping. Thus, DNN-based methods are potent for restoring clipped signals, even in severe conditions. However, since DNN has to learn the problem setting itself, DNN-based methods would not be able to directly use reliable samples for inference. Therefore, DNN-based methods cannot perform as well as sparse-optimization-based methods regarding restoration of slightly clipped signals.

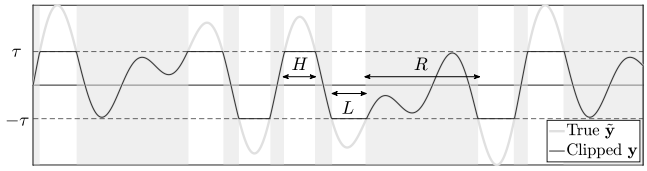

Fig. 1: Illustration of a clipped signal $\mathbf{y}$ with a clipping level τ (the solid black curve), and the corresponding true signal $\tilde{\mathbf{y}}$ (the solid gray curve). The areas with gray background indicate that clipping has no effect there. R is the set of the non-clipped indices, and H and L are those of the clipped indices above and below $\pm\tau$, respectively.

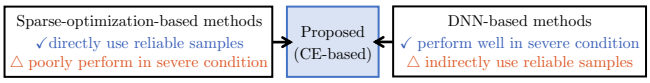

Fig. 2: Contrasting characteristics of conventional methods. The proposed method takes advantages of both kinds of methods.

In this paper, we propose a novel audio declipping method that fuses sparse-optimization-based and DNN-based methods and named it UPGLADE (**U**nplugged **P**lu**G**-and-p**L**ay **A**udio **DE**clipper, according to [14]). It utilizes *consensus equilibrium (CE)* [14], which is a framework for fusing multiple methods. CE can handle a wide range of methods (e.g., regularization and data fidelity) without the need for their mathematical expression [15–20]. By fusing sparse-optimization-based and DNN-based methods through CE, UPGLADE enjoys their advantages (see Fig. 2), compensating their disadvantages. Our experiments confirmed that the proposed method was more effective than conventional sparse-optimization-based and DNN-based methods in a wide range of clipping levels.

2. PRELIMINARIES

Let $\tilde{\mathbf{y}} \in \mathbb{R}^T$ be a digital audio signal whose length is T. Through *hard clipping* with a clipping level $\tau > 0$, the elements of $\tilde{\mathbf{y}}$ that exceed the dynamic range $[-\tau, \tau]$ are truncated as follows:

$$y[t] = \begin{cases} \tau & (\tilde{y}[t] \geq \tau) \\ \tilde{y}[t] & (-\tau < \tilde{y}[t] < \tau), \\ -\tau & (\tilde{y}[t] \leq -\tau) \end{cases} \tag{1}$$

where $y[t]$ is the t-th element of $\mathbf{y}$. The indices are split into three disjoint sets, $H = \{t \in [1,T] \,|\, y[t] \geq \tau\}$, $R = \{t \in [1,T] \,|\, |y[t]| < \tau\}$, and $L = \{t \in [1,T] \,|\, y[t] \leq -\tau\}$, as briefly shown in Fig. 1. The aim of audio declipping is to estimate the original signal $\tilde{\mathbf{y}}$ from the clipped signal $\mathbf{y}$ and the index information.

2.1. Sparse-optimization-based audio declipping

Sparse-optimization-based audio declipping methods [5–10] process a clipped signal by solving the following optimization problem:

$$\text{Find } \mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^T} S(\mathcal{G}\mathbf{x}) \text{ s.t. } \mathbf{x} \in \Gamma, \tag{2}$$

where S is a sparsity-promoting function (e.g., ℓ_0-norm [5]), $\mathcal{G}$ is the discrete Gabor transform (DGT) with a window $\mathbf{g} \in \mathbb{R}^T$,

$$(\mathcal{G}\mathbf{x})[m,n] = \sum_{t=1}^{T} x[t+an]\, g[t]\, e^{-2\pi \mathrm{i} mt/M}, \tag{3}$$

$\mathrm{i} = \sqrt{-1}$, $a \in \mathbb{N}$ denotes the time shifting step, $n \in [1, N]$ and $m \in [1, M]$ are the time and frequency indices, respectively, satisfying $aN = T$, and the indices are to be understood modulo T. The solution $\mathbf{x}^\star$ is naturally required to satisfy *clipping consistency* [8] derived from the problem setting Eq. (1). The time domain constraint in Eq. (2) ensures this consistency with the feasible set Γ,

$$\Gamma = \left\{ \mathbf{x} \in \mathbb{R}^T \;\middle|\; \begin{array}{ll} x[t] \geq \tau & (t \in H) \\ x[t] = y[t] & (t \in R) \\ x[t] \leq -\tau & (t \in L) \end{array} \right\}, \tag{4}$$

forcing the solution $\mathbf{x}^\star$ to be in Γ. However, since Γ has infinitely many signals, finding a solution from Γ is an ill-posed problem. In order to deal with this, *sparsity in the T-F domain* is utilized in Eq. (2), which is a classic regularization in audio declipping. Since clipping generates extra components in the T-F domain, minimizing $S(\mathcal{G}\mathbf{x})$ will remove them, and thus, the solution $\mathbf{x}^\star$ will be close to the original signal $\tilde{\mathbf{y}}$.

2.2. DNN-based audio declipping

DNN-based audio declipping methods have been studied lately. Still, compared to sparse-optimization-based methods, only a few DNN-based methods have been proposed [11–13]. DNN extracts relevant information to audio declipping from input features and is optimized to approximate a mapping that estimates the corresponding true signals from clipped signals. The input feature has been waveforms of clipped signals [13], spectrograms [12], and mel-frequency cepstrum coefficients [11]. The output also differs for each method (e.g., waveform, magnitude spectrogram [13], complex-valued filter [12], and mel-frequency cepstrum coefficients [11]).

3. PROPOSED METHOD: UPGLADE

In this paper, we propose a novel audio declipping method named UPGLADE. The proposed method is based on *consensus equilibrium (CE)* [14], which is a framework for merging multiple models [15–20]. UPGLADE is a fusion of two models based on sparse-optimization-based (see Sec. 2.1) and DNN-based (see Sec. 2.2) audio declipping through CE.

Sparse-optimization-based methods can preserve reliable samples considering the property of hard clipping, *clipping consistency*, via the time domain constraint in Eq. (2). Therefore, when an observed signal has abundant reliable samples (i.e., when clipping distortion is small), accurate restoration can be achieved. However, sparse-optimization-based methods might not work well in severe conditions (i.e., when clipping distortion is large). This is because sparse-optimization-based methods indirectly tackle audio declipping assuming that the true signals are sparse in the T-F domain.

DNN-based methods, on the other hand, learn the relationship between clipped signals and the corresponding true signals to directly understand and tackle audio declipping. This data-driven process results in effective restoration, even in severe conditions. However, when clipping distortion is small, DNN-based methods would not work as well as sparse-optimization-based methods. This is because DNN needs to learn even the clipping consistency itself. Adding to this, it is difficult for a DNN to express the identity map, which is deemed close to ideal when clipping distortion is small.

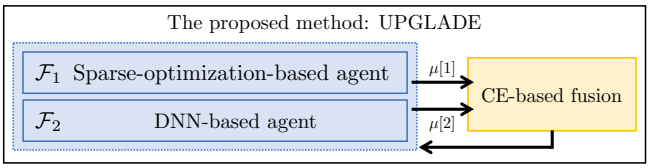

Fig. 3: Overview of our proposed method. Through CE-based fusion, sparse-optimization-based and DNN-based audio declipping are both performed repeatedly, and their balance is determined upon $\mu[1]$ and $\mu[2]$.

The two kinds of conventional methods have their own advantages and disadvantages as discussed above. In short, they have almost contrasting characteristics depending on clipping level: the sparse-optimization-based method is effective when the distortion is small, while the DNN-based method is effective when the distortion is large. Therefore, if it is possible to properly combine the two kinds of methods, audio declipping that takes advantages of both would be realized. This is the motivation of our proposal.

3.1. Consensus Equilibrium (CE)

Consensus equilibrium [14–20] is a scheme for fusing multiple models. Let us consider the following consensus optimization problem:

$$\text{Minimize} \quad \sum_{k=1}^{K} \mu[k]\, f_k(\mathbf{x}_k) \quad \text{s.t.} \;\; \forall k,\, \mathbf{x}_k = \mathbf{x}, \tag{5}$$

where $f_k : \mathbb{R}^T \to \mathbb{R} \cup \{+\infty\}$, $\mu[k] > 0$, and $\sum_{k=1}^{K} \mu[k] = 1$. To extend this consensus optimization problem to the CE framework, K vector-valued maps $\mathcal{F}_k : \mathbb{R}^T \to \mathbb{R}^T$ (called *agents*) are defined corresponding to the functions f_k. These agents are in *equilibrium* with $(\mathbf{x}^\star, \mathbf{u}^\star) \in \mathbb{R}^{T \times KT}$ that satisfies the following equations:

$$\mathcal{F}_k(\mathbf{x}^\star + \mathbf{u}_k^\star) = \mathbf{x}^\star \quad (k = 1, \ldots, K), \tag{6a}$$

$$\sum_{k=1}^{K} \mu[k]\, \mathbf{u}_k^\star = \mathbf{0}, \tag{6b}$$

where $\mathbf{u}^\star$ is obtained by stacking the vectors $\mathbf{u}_1^\star, \ldots, \mathbf{u}_K^\star$ (i.e., $\mathbf{u}^\star = [\mathbf{u}_1^{\star\mathsf{T}}, \ldots, \mathbf{u}_K^{\star\mathsf{T}}]^\mathsf{T}$), and $\mathbf{x}^\mathsf{T}$ is the transpose of $\mathbf{x}$. When we define each agent $\mathcal{F}_k$ as the proximity operator [21, 22] of the corresponding f_k, the set of the solutions to the CE equation Eq. (6) is exactly the same with that to the consensus optimization problem Eq. (5). Moreover, CE is of great worth when $\mathcal{F}_k$ is not a proximity operator and there is no underlying optimization problem. The agent $\mathcal{F}_k$ does not explicitly require the corresponding original function f_k, and thus, the CE framework can cover a wide range of models (e.g., regularization, data fidelity, and data-driven prior). That is, the CE framework allows us to arbitrarily choose some models and to fuse them in a balance determined by $\mu[k]$.

3.2. UPGLADE: Unplugged PluG-and-pLay Audio DEclipper

Here, we propose UPGLADE, which utilizes the CE framework to fuse sparse-optimization-based and DNN-based methods. We first introduce two agents used in UPGLADE, $\mathcal{F}_1$ (based on sparse-optimization-based audio declipping) and $\mathcal{F}_2$ (based on DNN-based audio declipping). An algorithm to tackle Eq. (6) is afterwards introduced. Fig. 3 shows the overview of UPGLADE.

Sparse-optimization-based agent: The first agent $\mathcal{F}_1$ is *an optimization algorithm itself that tackles the problem Eq.* (2). Alg. 1 is the primal-dual splitting algorithm [22] applied to Eq. (2), where $\mathcal{G}^*$ is the adjoint of $\mathcal{G}$, $\mathcal{P}_\Gamma$ is the projection operator onto Γ [6, 8], τ and σ are step size parameters, and $\alpha^{[i]}$ is a relaxation factor [22]. As in

Algorithm 1 Sparse-optimization-based agent $\mathcal{F}_1(\mathbf{x}_1)$

Initialization: $\mathbf{x}^{[1]} = \mathbf{x}_1$, $\mathbf{v}^{[1]} = \mathcal{G}\mathbf{x}^{[1]}$, $\kappa^{[1]}$
Output: $\mathbf{x}^{[I+1]}$
for $i = 1, \ldots, I$ **do**
  $\mathbf{p}^{[i]} = \mathcal{P}_\Gamma(\mathbf{x}^{[i]} - \tau\,\sigma\,\mathcal{G}^*\mathbf{v}^{[i]})$
  $\mathbf{r}^{[i]} = \mathbf{v}^{[i]} + \mathcal{G}(2\,\mathbf{p}^{[i]} - \mathbf{x}^{[i]})$
  $\mathbf{q}^{[i]} = \mathbf{r}^{[i]} - \mathcal{T}_{\kappa^{[i]}}(\mathbf{r}^{[i]})$
  $(\mathbf{x}^{[i+1]}, \mathbf{v}^{[i+1]}) = (1-\alpha^{[i]})\,(\mathbf{x}^{[i]}, \mathbf{v}^{[i]}) + \alpha^{[i]}\,(\mathbf{p}^{[i]}, \mathbf{q}^{[i]})$
  $\kappa^{[i+1]} = \kappa^{[i]} + \delta$
end for

the state-of-the-art sparse-optimization-based method, ASPADE [5], we use $\mathcal{T}_\kappa^{[i]}$ that remains only $\kappa^{[i]}$ largest elements intact and sets the others zero in each time segment for promoting sparsity. Moreover, we increase $\kappa^{[i]}$ by δ at each iteration referring to ASPADE. Note that an entire signal is processed at once, whereas ASPADE processes each window. Since $\mathcal{F}_1$ is able to promote sparsity in the T-F domain maintaining clipping consistency, UPGLADE also explicitly makes use of reliable samples and takes advantage of the general property of acoustic signals.

DNN-based agent: As used in conventional audio declipping [13], we use a DNN to estimate the original waveform from an input clipped waveform. However, since partially restored data are completely unknown to the DNN, merely letting $\mathcal{F}_2(\mathbf{x}) = \mathrm{DNN}(\mathbf{x})$ may lead to deterioration in restoration quality over iteration (See Fig. 3). Therefore, in this study, we use the proximity operator of the squared error function $\|\mathbf{x} - \mathbf{d}\|^2/2$ with $\mathbf{d} = \mathrm{DNN}(\mathbf{y})$,

$$\mathcal{F}_2(\mathbf{x}) = \lambda\,\mathbf{x} + (1-\lambda)\,\mathbf{d}, \tag{7}$$

instead, so that the output is in close proximity to $\mathbf{x}$, and still retains the effect of the DNN. $\lambda \in [0, 1]$ is used to adjust how much $\mathbf{d}$ is considered. Since $\mathcal{F}_2$ utilizes the inference result of a DNN, UPGLADE is able to directly understand and tackle audio declipping.

Fixed point algorithm: Alg. 2 shows the resulting algorithm of CE that uses the two agents $\mathcal{F}_1$ and $\mathcal{F}_2$, where $\rho > 0$ is a step size parameter. The CE equation Eq. (6) for the proposed method can be reformulated as the simple equation as follows:

$$\mathcal{F}(\mathbf{z}^\star) = \mathcal{A}_{\boldsymbol{\mu}}(\mathbf{z}^\star), \tag{8}$$

where $\mathbf{z}_k^\star = \mathbf{x}^\star + \mathbf{u}_k^\star$, $\mathbf{z}^\star = [\mathbf{z}_1^{\star\mathsf{T}}, \mathbf{z}_2^{\star\mathsf{T}}]^\mathsf{T}$, $\bar{\mathbf{z}}_{\boldsymbol{\mu}}^\star = \mu[1]\,\mathbf{z}_1^\star + \mu[2]\,\mathbf{z}_2^\star$, $\mathcal{F}(\mathbf{z}^\star) = [\mathcal{F}_1(\mathbf{z}_1^\star)^\mathsf{T}, \mathcal{F}_2(\mathbf{z}_2^\star)^\mathsf{T}]^\mathsf{T}$, and $\mathcal{A}_{\boldsymbol{\mu}}(\mathbf{z}^\star) = [\bar{\mathbf{z}}_{\boldsymbol{\mu}}^{\star\mathsf{T}}, \bar{\mathbf{z}}_{\boldsymbol{\mu}}^{\star\mathsf{T}}]^\mathsf{T}$. Moreover, due to the linearity of $\mathcal{A}_{\boldsymbol{\mu}}$, Eq. (8) can be further reformulated into the following fixed point problem using the identity operator Id:

$$\text{Find}\ \ \mathbf{z}^\star\ \ \text{s.t.}\ \ (2\,\mathcal{F} - \mathrm{Id})(2\,\mathcal{A}_{\boldsymbol{\mu}} - \mathrm{Id})(\mathbf{z}^\star) = \mathbf{z}^\star. \tag{9}$$

Alg. 2 tackles this problem by using the Mann iteration [23]. Although the convergence of Alg. 2 cannot be guaranteed, we will experimentally see its behavior in Sec. 4.3.

4. EXPERIMENTS AND RESULTS

4.1. Training

LIBRI speech corpus [24] was used for training. We used 28000 clean speech signals for training, and 16384 samples of a voiced part were extracted from each signal (about 1 s at sampling frequency of 16 kHz). All data were peak-normalized, scaling all samples to have a maximum absolute value of 1. They were corrupted by clipping with a clipping level τ randomly drawn from the uniform distribution in the interval $[0.01, 0.99]$.

Algorithm 2 Proposed method: UPGLADE

Input: ${\mathbf{z}_1}^{[1]} = {\mathbf{z}_2}^{[1]} = \mathcal{P}_\Gamma(\mathbf{d})$
Output: $\mathcal{P}_\Gamma(\mu[1]\,{\mathbf{z}_1}^{[J+1]} + \mu[2]\,{\mathbf{z}_2}^{[J+1]})$
for $j = 1, \ldots, J$ **do**
  $\tilde{\mathbf{z}}_1^{[j]} = 2\,\mathcal{F}_1({\mathbf{z}_1}^{[j]}) - {\mathbf{z}_1}^{[j]}$
  $\tilde{\mathbf{z}}_2^{[j]} = 2\,\mathcal{F}_2({\mathbf{z}_2}^{[j]}) - {\mathbf{z}_2}^{[j]}$
  $\tilde{\mathbf{z}}^{[j]} = \mu[1]\,\tilde{\mathbf{z}}_1^{[j]} + \mu[2]\,\tilde{\mathbf{z}}_2^{[j]}$
  ${\mathbf{z}_1}^{[j+1]} = (1-\rho)\,{\mathbf{z}_1}^{[j]} + \rho\,(2\,\tilde{\mathbf{z}}^{[j]} - \tilde{\mathbf{z}}_1^{[j]})$
  ${\mathbf{z}_2}^{[j+1]} = (1-\rho)\,{\mathbf{z}_2}^{[j]} + \rho\,(2\,\tilde{\mathbf{z}}^{[j]} - \tilde{\mathbf{z}}_2^{[j]})$
end for

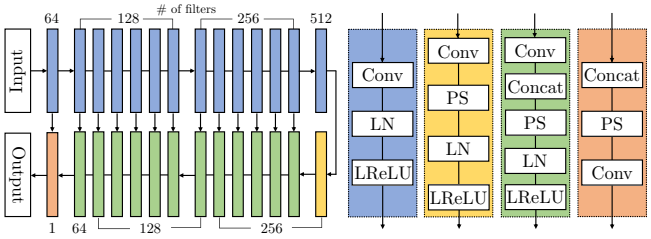

Fig. 4: Network structure. "Conv", "LN", "LReLU", "Concat", and "PS" stand for convolution, layer normalization, leaky ReLU, concatenation, and pixel shuffle, respectively. Conv was performed with a kernel of width 5 and stride 2. The scale of LReLU was 0.3.

The structure of the DNN for the second agent $\mathcal{F}_2$ is shown in Fig. 4. It was designed with reference to T-UNet [13], but it uses layer normalization instead of batch normalization for adaptability to small batch sizes. Adding to this, we set the bias of the last Conv to zero to prevent biasing the output time waveform. See Fig. 4 and its caption for the parameters of each layer. The DNN was trained 100 epochs with the adam optimizer [25] with a batch size of 4, a learning rate of 0.0001, and decay rates of $\beta_1 = 0.5$ and $\beta_2 = 0.9$. The loss function was the time-domain mean-squared-error (MSE).

4.2. Testing

For testing, 200 speech signals [26] sampled at 16 kHz from the TIMIT corpus were used. The clipping level τ was set such that signal-to-distortion ratio (SDR), $20\,\log_{10}\|\mathbf{y}\|_2/\|\mathbf{y} - \tilde{\mathbf{y}}\|_2$, was any one of 1, 3, 5, 10, and 15 dB. All the data were cut out into 16384 samples and peak-normalized. 75 % overlapped 1024-point-long-Hann window was used for DGT. The parameters were as follows: the default value of $\boldsymbol{\mu} = [\mu[1], \mu[2]]^\mathsf{T}$ was $[1, 1]^\mathsf{T}/2$, $\rho = 0.1$, λ was linearly increased from 0.5 to 1, $\tau = \sigma = \alpha^{[i]} = 1$, $\kappa^{[1]} = M/4$, and $\delta = 1$. The maximum iteration counts I and J were 200 and 20, respectively. We used ΔSDR and ΔPESQ, which are improvement of SDR and PESQ [27], respectively, for evaluation.

Some conventional audio declipping methods were also performed for comparison. As the state-of-the-art sparse-optimization-based methods, ASPADE [5] and Parabola-Weighted ℓ_1 minimization (PWℓ_1) [6] were performed. APPLADE [28] was performed as a precedent combining DNN and sparse optimization by the plug-and-play (PnP) framework [29, 30]. For comparison to DNN-based audio declipping, inference results of the DNN (Fig. 4) were used. Moreover, as simple ways to combine sparse optimization (ASPADE as a representative) and DNN (Fig. 4 as a representative), the results of their series connection (referring to *ASPADE first* as A→D and *DNN first* as D→A) were also compared. The parameters and other detailed settings were taken from the original paper of each or [8] to suit for the speech signals sampled at 16 kHz.

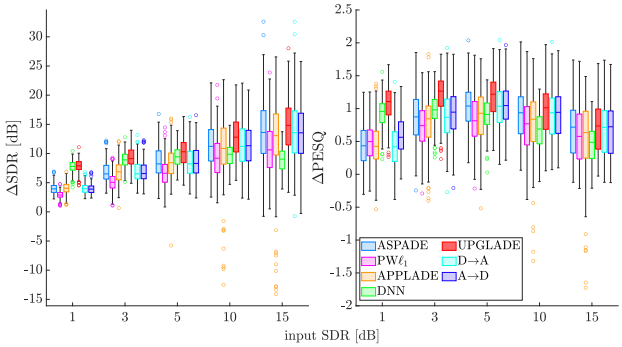

Fig. 5: Overall comparison among the conventional methods, the series connections, and UPGLADE.

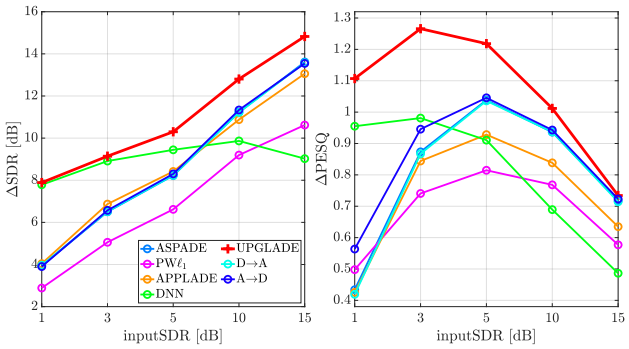

Fig. 6: Comparison in terms of the median of ΔSDR and ΔPESQ.

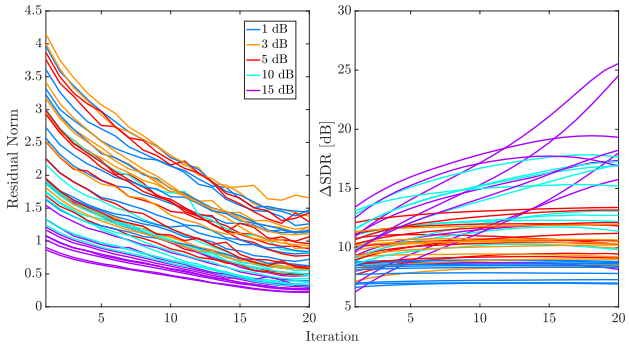

Fig. 7: Behavior of UPGLADE over iterations in terms of the residual norm and ΔSNR. The legend indicates input SDR.

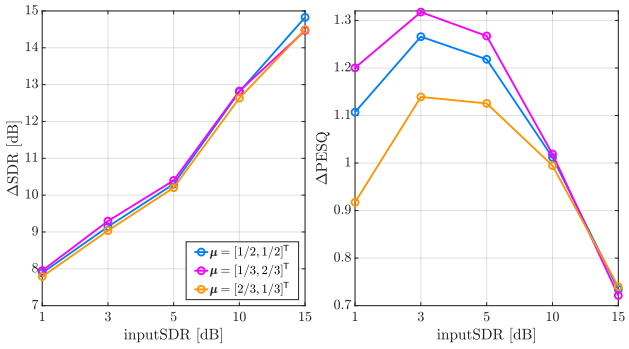

Fig. 8: Behavior of UPGLADE at different values of $\boldsymbol{\mu}$.

4.3. Results

The overall comparison is shown in Fig. 5, and only the median values are shown in Fig. 6 for ease of viewing. First, the sparse-optimization-based methods (ASPADE and PWℓ_1) had relatively high ΔSDR at high input SDR (i.e., small distortion). However, their ΔSDR at low input SDR was nearly 10 dB lower than that at high input SDR. Similarly, ΔPESQ was poor at low input SDR. In contrast, ΔSDR and ΔPESQ of DNN-based restoration were considerably higher than those of the sparse-optimization-based methods at low input SDR. However, DNN-based restoration was inferior to them at high input SDR.

Next, we discuss the results of the methods that combine sparse-optimization-based and DNN-based approaches. APPLADE, inspired by PnP, showed an improvement in ΔSDR from ASPADE, especially at low input SDR. However, it was unstable with a large variation in values (See Fig. 5). Furthermore, ΔPESQ significantly dropped compared to ASPADE. As for the methods that connect ASPADE and DNN in series (A→D and D→A), the results were not much different from those of ASPADE in both cases. This may be because: in the former, the restoration results by ASPADE were unknown to the DNN, and so the DNN did not work well; and in the latter, ASPADE was to some extent robust to the initial value, and so using the output of the DNN as the initial value did not have much effect. In conclusion, it was found that connecting the two methods in series did not improve audio declipping performance. The proposed method, UPGLADE, achieved the same or better performance than all the other methods at all the input SDR. Furthermore, it had less scattered values as APPLADE did. Therefore, the effectiveness of the proposed method was confirmed.

To discuss the stability of the proposed method, *residual norm* and ΔSDR at each iteration for 50 randomly selected data are shown in Fig. 7, where the legend indicates the input SDR. The residual norm is an evaluation measure for the convergence of CE [14], and is expressed by the following formula: $\|\mathcal{F}(\mathbf{z}^{[k]}) - \mathcal{A}_{\boldsymbol{\mu}}(\mathbf{z}^{[k]})\|$. According to it, UPGLADE enjoyed stable convergence to some extent. In addition, ΔSDR improved from the initial value (which is $\mathcal{P}_\Gamma(\mathbf{d})$, i.e., the DNN restoration result) in many cases. The improvement was especially remarkable at high input SDR, and the residual norm at high input SDR tended to be lower than other input SDR. Further improvement in stability and performance may be expected by utilizing the residual norm for a stopping criterion of the algorithm.

Finally, the median values of ΔSDR and ΔPESQ at different values of $\boldsymbol{\mu}$, which determines the balance of CE, are shown in Fig. 8. The blue color indicates $\boldsymbol{\mu} = [1,1]^\mathsf{T}/2$, which is the same with the results of UPGLADE in Fig. 6. In comparison with $\boldsymbol{\mu} = [2,1]^\mathsf{T}/3$ and $\boldsymbol{\mu} = [1,2]^\mathsf{T}/3$, there was little difference in ΔSDR, but a notable difference in ΔPESQ: the more valued $\mathcal{F}_2$ was, the higher ΔPESQ became, especially at low input SDR. Therefore, the results of UPGLADE in Fig. 6 were not optimal, and further performance improvement can be expected by adjusting $\boldsymbol{\mu}$.

5. CONCLUSIONS

In this paper, we proposed the CE-based audio declipping, named UPGLADE. Both sparse-optimization-based and DNN-based audio declipping methods are utilized in UPGLADE, and thus, it enjoys their advantages, compensating their disadvantages. Our experiments showed that UPGLADE was effective in a wide range of clipping levels, unlike the conventional methods. Moreover, it was suggested that utilizing the residual norm and adjusting $\boldsymbol{\mu}$ could further improve the performance of the proposed method. Future work will be on theoretical guarantees of convergence and applying to other audio restoration problems [31, 32].

6. REFERENCES

[1] C.-T. Tan and B.C.J. Moore, "Perception of nonlinear distortion by hearing-impaired people," *Int. J. Audiol.*, vol. 47, no. 5, pp. 246–256, 2008.

[2] K.H. Arehart, J.M. Kates, and M.C. Anderson, "Effects of noise, nonlinear processing, and linear filtering on perceived music quality," *Int. J. Audiol.*, vol. 50, no. 3, pp. 177–190, 2011.

[3] J. Malek, "Blind compensation of memoryless nonlinear distortions in sparse signals," in *21st Eur. Signal Process. Conf. (EUSIPCO)*, 2013, pp. 1–5.

[4] Y. Tachioka, T. Narita, and J. Ishii, "Speech recognition performance estimation for clipped speech based on objective measures," *Acoust. Sci. Technol.*, vol. 35, no. 6, pp. 324–326, 2014.

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

[6] P. Záviška, P. Rajmic, and J. Schimmel, "Psychoacoustically motivated audio declipping based on weighted ℓ_1 minimization," in *Int. Conf. Telecommun. Signal Process. (TSP)*, 2019, pp. 338–342.

[7] S. Emura and N. Harada, "An extension of sparse audio declipper to multiple measurement vectors," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 686–690.

[8] C. Gaultier, S. Kitić, R. Gribonval, and N. Bertin, "Sparsity-based audio declipping methods: Selected overview, new algorithms, and large-scale evaluation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1174–1187, 2021.

[9] P. Záviška, P. Rajmic, Al. Ozerov, and L. Rencker, "A survey and an extensive evaluation of popular audio declipping methods," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 1, pp. 5–24, 2021.

[10] P. Záviška and P. Rajmic, "Audio declipping with (weighted) analysis social sparsity," in *2022 45th Int. Conf. Telecommun. Signal Process. (TSP)*, 2022, pp. 407–412.

[11] F. Bie, D. Wang, J. Wang, and T.F. Zheng, "Detection and reconstruction of clipped speech in speaker recognition," *Speech Commun.*, vol. 72, 07 2015.

[12] W. Mack and E.A.P. Habets, "Declipping speech using deep filtering," in *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2019, pp. 200–204.

[13] A.A. Nair and K. Koishida, "Cascaded time + time-frequency Unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 7153–7157.

[14] G.T. Buzzard, S.H. Chan, S. Sreehari, and C.A. Bouman, "Plug-and-Play unplugged: Optimization-free reconstruction using consensus equilibrium," *SIAM J. Imaging Sci.*, vol. 11, no. 3, pp. 2001–2020, 2018.

[15] P. Goyes-Peñafiel, E. Vargas, C.V. Correa, W. Agudelo, B. Wohlberg, and H. Arguello, "A consensus equilibrium approach for 3-D land seismic shots recovery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] M. Hossain, S.C. Paulson, H. Liao, W.W. Chen, and C.A. Bouman, "Ultra-sparse view reconstruction for flash X-ray imaging using consensus equilibrium," *2020 54th Asilomar Conf. Signals Syst. Comput.*, pp. 631–635, 2020.

[17] M.U. Ghani and W.C. Karl, "Data and image prior integration for image reconstruction using consensus equilibrium," *IEEE Trans. Comput. Imaging*, vol. 7, pp. 297–308, 2021.

[18] R. Hyder, H. Mansour, Y. Ma, P.T. Boufounos, and P. Wang, "A consensus equilibrium solution for deep image prior powered by RED," in *2021 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 1380–1384.

[19] Q. Zhai, B. Wohlberg, G.T. Buzzard, and C.A. Bouman, "Projected multi-agent consensus equilibrium for ptychographic image reconstruction," in *2021 55th Asilomar Conf. Signals Syst. Comput.*, 2021, pp. 1694–1698.

[20] V. Sridhar, X. Wang, G.T. Buzzard, and C.A. Bouman, "Distributed iterative CT reconstruction using multi-agent consensus equilibrium," *IEEE Trans. Comput. Imaging*, vol. 6, pp. 1153–1166, 2020.

[21] N. Parikh and S. Boyd, *Proximal Algorithms*, Now Publishers Inc., 2014.

[22] N. Komodakis and J.-C. Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 31–54, 2015.

[23] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer Publ. Co. Inc., 1st edition, 2011.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 5206–5210.

[25] D.P. Kingma and J.L. Ba, "Adam: A method for stochastic optimization," in *Proc. IEEE Int. Conf. on Learn. Represent. (ICLR)*, 2015.

[26] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*, Hoboken, NJ, USA: Wiley, 11 2016.

[27] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2001, vol. 2, pp. 749–752.

[28] T. Tanaka, K. Yatabe, M. Yasuda, and Y. Oikawa, "APPLADE: Adjustable plug-and-play audio declipper combining dnn with sparse optimization," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 1011–1015.

[29] S.V. Venkatakrishnan, C.A. Bouman, and B. Wohlberg, "Plug-and-Play priors for model based reconstruction," in *IEEE Glob. Conf. Signal Inf. Process.*, 2013, pp. 945–948.

[30] S.H. Chan, X. Wang, and O.A. Elgendy, "Plug-and-Play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 84–98, 2017.

[31] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "VoiceFixer: Toward general speech restoration with neural vocoder," *arXiv:2109.13731*, 2021.

[32] J. Serrá, S. Pascual, J. Pons, R.O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv:2206.03065*, 2022.