時間周波数領域におけるグリッド間の整合性に基づくクリッピングの除去*

☆三浦晋, 宮部滋樹, 山田武志, 牧野昭二 (筑波大),
中島弘史 (工学院大), 中臺一博 (HRI)

1 はじめに

信号を高品質で録音するためには、入力ゲインを適切に設定して量子化ノイズを減らすべきである。しかし、入力ゲインが高すぎる場合は録音した信号の振幅レベルが限界を超えてひずんでしまうことがある。信号の振幅が、録音機器で扱える振幅の範囲を超えたとき、限界のレベル以上の値は限界のレベルの値で切り揃えられてしまい、波形は Fig. 1 のように変形してしまう。この現象はクリッピングと呼ばれ、音質の劣化だけでなく、音声認識、音源分離、方向推定などの性能を低下させる原因の一つとなっている。

クリッピングされた信号は原信号の情報を失っているため、完全な修復は非常に困難である。しかしながら、修復対象とする信号から得られる事前知識を用いて原信号の振幅をある程度推定することは可能である。対象信号の事前知識を用いる従来手法として、対象信号が帯域制限されている場合にオーバーサンプリングの性質を利用して修復する方法 [1]、音声信号を対象とした修復方法 [2][3]、統計的な仮定を用いる方法 [4] などが提案されている。対象信号の事前知識を用いずにクリッピング修復を行うことはさらに難しい問題となる。クリッピングの前後区間の波形代入を用いたいくつかの試みがなされているが [5]、元々存在しなかった高周波ノイズが新たに発生してしまうといった問題が確認されている。本研究ではこれらのノイズを防ぎつつ修復を行うため、時間周波数領域でのグリッド間の整合性を考慮して修復を行う手法を提案し、従来手法との性能比較を行う。

2 従来手法

2.1 RVP を用いたクリッピング修復 [5][6]

我々はこれまでに RVP(逐次ベクトル射影法)[7] を用いたクリッピング修復手法を提案した。RVP とは、対象とする信号と相関が高い基底信号を一定の本数選択し、基底信号の線形和によって修復信号を表現する。全ての基底を用いないことにより、対象とする信号から不要な基底信号が除かれることをねらいとしている。基底信号としてフーリエ基底を用い、RVP の相関の計算にクリッピングしていない部分のみを用いることにより、クリッピングしていない部分においてスパースな表現で近似された信号が得られる。このようにして得られた近似信号を、クリッピング部分に延長することにより、クリッピングしていない部分との整合性を保った推定信号が得られる。この手法ではクリッピング部分を検出して、その直前と直後の一定区間を用いて修復を行なう。修復性能は SNR にして 2 dB 程度回復させられるが、高周波数域に新たなノイズが重畳してしまうという課題がある。

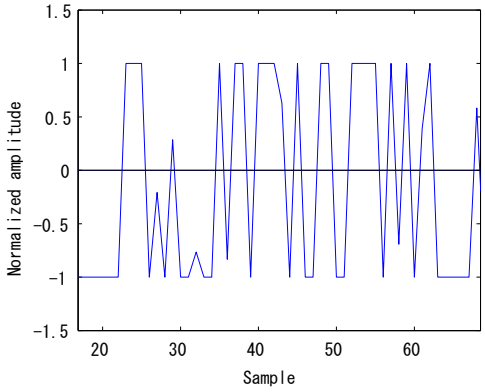

Fig. 1 An example of clipped signal.

2.2 Adler らの手法 [8]

Adler らの手法 [8] は、クリッピングしていない部分の相関を用いた RVP と等価な基底表現と、基底信号の振幅に関する拘束条件を用いたクリッピング修復法で、上述の手法と同時期に発表されている。RVP を用いたクリッピング修復との相違点は以下の 2 点である。クリッピングの位置によらず、フレーム長が固定のフレーム分析を行っている点、クリッピング部分の修復波形の振幅に適切な上限と下限を設定している点である。性能評価実験では、様々な音声・楽音・クリッピングレベルに対して SNR にして 4–5 dB 程度の改善が得られることが示されている。

3 提案手法

3.1 提案手法のアプローチ

従来法では時間領域での整合性のみを考慮していたため、周波数領域での整合性が保たれていなかった。これにより新たな雑音が重畳しているのではないかと仮定される。提案手法では従来法で考慮されていなかった周波数領域での整合性について考慮することにより性能の向上を図る。

分析対象として時間領域での波形でなく、スペクトログラムを扱うことで実現を試みる。また、本手法でも従来法と同様にクリッピングしている部分をスパースな表現により近似するが、ここで用いるフーリエ基底を選択する際に、以下のような基準を設けることで周波数領域での整合性を考慮した修復を試みる。

基準 1 クリッピングされているサンプル数が多いフレームほど修復の程度を大きくする。

基準 2 クリッピングによる影響が大きい高周波数帯域ほど修復の程度を大きくする。

---

*Restoration of clipped acoustic signal based on the consistency among time-frequency grids, Shin Miura, Shigeki Miyabe, Takeshi Yamada, Shoji Makino(University of Tsukuba), Hirofumi Nakajima(Kogakuin University), Kazuhiro Nakadai(HRI)

基準 3 クリッピングによって発生するノイズに多く含まれる高周波成分の基底信号の使用を抑える。

3.2 周波数領域へのマッピング

対象信号のスペクトログラムを扱うため、対象信号を周波数領域へマッピングする。周波数領域へのマッピングは短時間フーリエ変換 (STFT) により行う。入力信号 $y(t)$ を STFT をした後、絶対値をとることで得られたスペクトログラムを N 行 K 列の行列 $F(n,k)$ とする。ここで、N は $F(n,k)$ の総フレーム数、K は $F(n,k)$ の周波数ビン数である。また、本手法では対象信号のスペクトログラムのうち N フレームずつを用いて分析対象区間とする。ここで、以降の処理の簡略化の為にスペクトログラムをベクトル化しておく。ここで得られたスペクトログラムを $F(n,k)$ とすると、ベクトル化されたスペクトログラム $\boldsymbol{f}$ は

$$\begin{aligned}\boldsymbol{f} &= [f_1,\cdots,f_M]^T \\ &= [F(1,0),\cdots,F(N,0), \\ &\quad F(1,1),\cdots,F(N,1),\cdots, \\ &\quad F(1,K-1),\cdots,F(N,K-1)]^T \quad (1)\\ M &= NK \quad (2)\end{aligned}$$

となる。

3.3 スペクトログラム推定アルゴリズム

推定スペクトログラムは RVP を用いて選択された基底信号の線形和で得られる。基底信号の選択順は、3.1 節での基準を考慮したペナルティ J_i を用いて決定する。$\boldsymbol{x}_i$ を正規形の二次元フーリエ基底、$\boldsymbol{f}$ を対象スペクトログラムとすると、J_i は以下の式で表される。

$$J_i = ||\boldsymbol{W}\boldsymbol{f} - \boldsymbol{W}\boldsymbol{x}_i a_i||^2 + \lambda |h_i a_i|^2 \quad (3)$$

ここで、$\boldsymbol{W}$ は $\boldsymbol{f}$ の各要素の基底表現による誤差を重みづける対角行列、h_i は基底重みを表す。これらについては後述する。λ は誤差重みと基底重みのどちらを重視するかを調整するパラメータである。また、係数 a_i は以下の偏微分

$$\frac{\partial J_i}{\partial a_i} = -2(\boldsymbol{W}\boldsymbol{f})^T(\boldsymbol{W}\boldsymbol{x}_i) + 2a_i||\boldsymbol{W}\boldsymbol{x}_i||^2 + 2\lambda {h_i}^2 a_i \quad (4)$$

を 0 とおくことにより、

$$a_i = \frac{(\boldsymbol{W}\boldsymbol{f})^T(\boldsymbol{W}\boldsymbol{x}_i)}{||\boldsymbol{W}\boldsymbol{x}_i||^2 + \lambda h_i^2} \quad (5)$$

と得られ、これを式 (3) に代入することで

$$J_i = -\frac{((\boldsymbol{W}\boldsymbol{f})^T(\boldsymbol{W}\boldsymbol{x}_i))^2}{||\boldsymbol{W}\boldsymbol{x}_i||^2 + \lambda h_i^2} \quad (6)$$

が得られる。このペナルティ J_i の値が小さいものほど基底信号として選択されやすい。そして、得られた式 (6) を最小化する $\boldsymbol{x}_i$ を選択する。この際、$\boldsymbol{f}$ は以下の式で更新される。

$$\boldsymbol{f} \leftarrow \boldsymbol{f} - a_m \boldsymbol{x}_m \quad (7)$$

ここで、m は選択された基底信号のインデックスである。以上の演算を選択された基底信号の本数が上限に達するまで繰り返す。全ての基底信号を選択し終えた後、選択された基底信号と係数の線形和 $\hat{\boldsymbol{f}} = \sum_i a_i \boldsymbol{x}_i$ をもって推定波形とする。

3.4 基底信号の設計

式 (3) で用いられている基底信号 $\mathbf{x}_i$ には、時間・周波数間で連続で整合のとれた信号を推定するため、2 次元コサイン基底を加工したものを用いる。基準 3 を満たすため、低域に大きなパワーをもつよう適当に設定したパラメタ d を用いて、$(1-k/K)^d$ でコサイン基底を重みづけたものを基底信号として用いる。基底信号 $\mathbf{x}_i$ は以下の式であらわされる。

$$\begin{aligned}\mathbf{x}_i &= [x_1,\cdots,x_M]^T \\ &= \big[X_i(0,0),\cdots,X_i(N-1,0), \\ &\quad X_i(0,1),\cdots,X_i(N-1,1),\cdots, \\ &\quad X_i(0,K-1),\cdots,X_i(N-1,K-1)\big]^T \quad (8)\end{aligned}$$

$$X_{uK+v}(n,k) = \begin{cases} \left(\frac{K-k+1}{K}\right)^d \sin\left(2\pi\frac{uk}{K}\right)\sin\left(2\pi\frac{vn}{N}\right) \\ \qquad \left(0 \le u < \frac{K}{2}, 0 \le v < \frac{K}{2}\right) \\ \left(\frac{K-k+1}{K}\right)^d \sin\left(2\pi\frac{uk}{K}\right)\cos\left(2\pi\frac{vn}{N}\right) \\ \qquad \left(0 \le u < \frac{K}{2}, \frac{N}{2} \le v < N\right) \\ \left(\frac{k+1}{K}\right)^d \cos\left(2\pi\frac{uk}{K}\right)\sin\left(2\pi\frac{vn}{N}\right) \\ \qquad \left(\frac{K}{2} \le u < K, 0 \le v < \frac{N}{2}\right) \\ \left(\frac{k+1}{K}\right)^d \cos\left(2\pi\frac{uk}{K}\right)\cos\left(2\pi\frac{vn}{N}\right) \\ \qquad \left(\frac{K}{2} \le u < K, \frac{N}{2} \le v < N\right) \end{cases} \quad (9)$$

3.5 各重みの計算

3.1 節での 3 つの基準を考慮するために設定した重みについて説明する。

3.5.1 誤差重み

式 (3) の第一項は対象スペクトログラムと推定スペクトログラムの二乗誤差を表しており、ここでの誤差重み $\boldsymbol{W}$ は、3.1 節の基準 1 と 2 の役割を果たす。誤差重みはクリッピング重み $\boldsymbol{w}_c$ と帯域重み $\boldsymbol{w}_f$ によって求められる。まず、基準 1 を表す重みであるクリッピング重み $\boldsymbol{w}_c$ について説明する。この重みの n 番目の要素は n フレーム目におけるクリッピングされていない時間の割合を表す。値が小さいほどクリッピングによる影響が大きいフレームとされる。クリッピング重みベクトル $\boldsymbol{w}_c$ と、その各要素を以下の式で定義する。

$$\begin{aligned}\boldsymbol{w}_c &= [w_c(1), w_c(2),\cdots,w_c(N)]^T \quad (10)\\ w_c(n) &= 1 - \frac{L_c(n)}{L} \quad (11)\end{aligned}$$

ここで、n はフレーム番号、N は総フレーム数、L はフレームの時間長、$L_c(n)$ は n フレーム目におけるクリッピングを受けている時間である。

次に、基準 2 を表す重みである帯域重みについて説明する。この重みの k 番目の要素は k 番目の周波数ビンの信頼度を表している。クリッピングにより高周波数域が劣化しやすいため、高周波数域での信頼度が低くなるように設定する。帯域重みベクトル $\boldsymbol{w}_f$ は以下の計算式で表現した。

$$\begin{aligned}\boldsymbol{w}_f &= [w_f(0), w_f(1),\cdots,w_f(K-1)]^T \quad (12)\\ w_f(k) &= \frac{1}{k} \quad (13)\end{aligned}$$

ここで、k は周波数ビン番号である。これは自然な音として聴くことができるような性質を持つピンクノイズのスペクトル構造に準拠したものである。そして、誤差重み $\boldsymbol{w}$ と誤差重み行列 $\boldsymbol{W}$ は N 行 K 列の行列 $\boldsymbol{W}_0$ を用いて以下のように得られる。

$$\begin{aligned}\boldsymbol{W}_0 &= \boldsymbol{w}_f \boldsymbol{w}_c && (14)\\ \boldsymbol{w} &= [W_0(1,0),\cdots,W_0(N,0),\\ &\quad W_0(1,1),\cdots,W_0(N,1),\cdots,\\ &\quad W_0(1,K-1),\cdots,W_0(N,K-1)]^T && (15)\\ &\quad \boldsymbol{W} = diag(\boldsymbol{w}_c \boldsymbol{w}_f^T) && (16)\end{aligned}$$

3.5.2 基底重み

式 (3) の第二項は使用する基底信号の重み付き二乗ノルムを表しており、これに乗ずる基底重み h_i は基準 3 の役割を果たす。高周波数の基底信号でペナルティ J_i の値が大きくなるように、高周波数ビンでは値が高くなるように設定した。h_i は $\boldsymbol{w}_f$ を用いて以下の式で表される

$$\begin{aligned}\boldsymbol{h} &= [h_1,\cdots,h_M]\\ &= [w_f(K),w_f(K-1),\cdots,w_f(1),\\ &\quad w_f(K),w_f(K-1),\cdots,w_f(1),\cdots,\\ &\quad w_f(K),w_f(K-1),\cdots,w_f(1)]^T \qquad (17)\end{aligned}$$

4 後処理

入力信号はクリッピングの影響により原信号よりもパワーが小さく観測されるため、入力信号より推定された推定スペクトログラムもまたパワーが小さく推定されてしまう。この影響を緩和する後処理として、クリッピングによるパワーの減少を前もって見積もり、それを補償するよう推定信号を増幅する。

クリッピングして観測されるサンプルの、クリッピングの影響を受けない本来の平均振幅を前もって p として見積もると、分析範囲のパワー補償ゲイン P は以下のように与えられる。

$$P = \frac{\sum y'(t)^2}{\sum y(t)^2} \qquad (18)$$

$$y'(t) = \begin{cases} p & \text{if } y(t) \text{ is clipped} \\ y(t) & \text{otherwise} \end{cases} \qquad (19)$$

ここで $y(t)$ はクリッピングを受けて観測される入力信号を表す。RVP により推定された振幅スペクトログラムを $\mathbf{Y}$ とすると、パワーを保証された振幅スペクトログラム $\hat{\mathbf{Y}}$ は以下のように求められる。

$$\hat{\mathbf{Y}} = \sqrt{P}\mathbf{Y} \qquad (20)$$

パワーを補償された振幅スペクトログラム $\hat{\mathbf{Y}}$ を時間領域に変換して修復波形とする。

5 修復実験

提案手法の性能を評価するために、クリッピングを起こした信号の修復実験を行った。意図的にクリッピングした音声信号や楽曲信号をいくつか用意し、修復を施す前後での差異を比較することで評価を行った。各パラメータは手動により設定した。それぞれの値は Table. 2 の通りである。ここでのクリッピングレベルとは、信号をクリッピングする際の振幅の上限である。尚、各信号は最大振幅が 1 になるように正規化されている。

Table 1 Parameters

FFT 点数	512 点
オーバーラップ	1/2
窓関数	ハニング窓
クリッピングレベル	0.1, 0.2, 0.3, 0.4, 0.5
基底信号選択本数	100 本
λ	1024
N	4
p	0.5
d	0, 2

5.1 評価に用いたデータ

本評価実験では音声 20 種類と楽曲 5 種類のデータを用いた。音源は 2008 Signal Separation Evaluation Campaign のものを使用した。音源のサンプリング周波数は音声音楽共に 16 kHz、量子化ビット数は 16 である。

5.2 評価尺度

評価尺度は SNR を用いた。SNR は次式で表され、値が大きいほど音質が良いことを表す。

$$SNR(dB) = 10\log\frac{\Sigma y(t)^2}{\Sigma(y(t)-\hat{y}(t))^2} \qquad (21)$$

ここで、$y(t)$ はクリッピング前の信号、$\hat{f}(t)$ は修復後の信号である。どちらもクリッピング区間のみを切り出している。

5.3 実験結果・考察

Adler らの手法との音声信号についての SNR 比較を Fig. 2 と Fig. 3 に示す。また、本手法適用後の波形の例を Fig. 4 に、スペクトログラムの例を Fig. 5 に示す。Fig. 2 と Fig. 3 より、音源の種類やクリッピングレベルにより差はあるものの、従来手法よりも平均的な修復性能が向上していることが分かる。また、Fig. 5 より、周波数領域で見てもクリッピングによるノイズがある程度除去されていることが分かる。

6 まとめ

本研究では、RVP での基底信号選択時に各種重みを用いたペナルティ関数を用いることで周波数領域でのグリッド間の整合性を考慮し修復を行うクリッピング修復アルゴリズムを提案した。修復実験の結果、SNR 改善量において従来法を上回る性能が確認された。

参考文献

[1] J. S. Abel and J. O. Smith, “Restoring a clipped signal,” Proc. ICASSP, pp.1745–1748, 1991.

[2] A. Janssen, R. Veldhuis, and L. Vries, “Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes,”

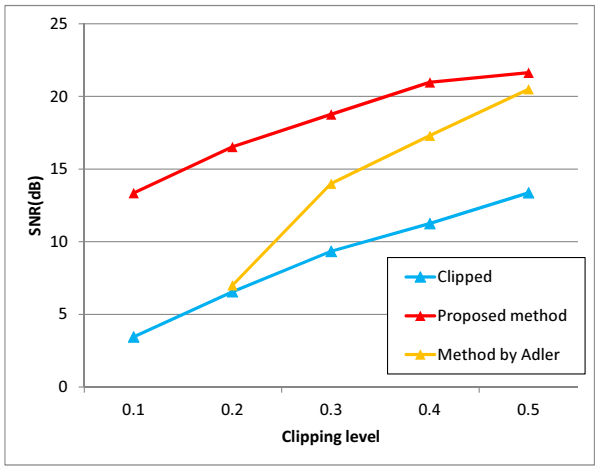

Fig. 2 SNR performance of the proposed method (speech).

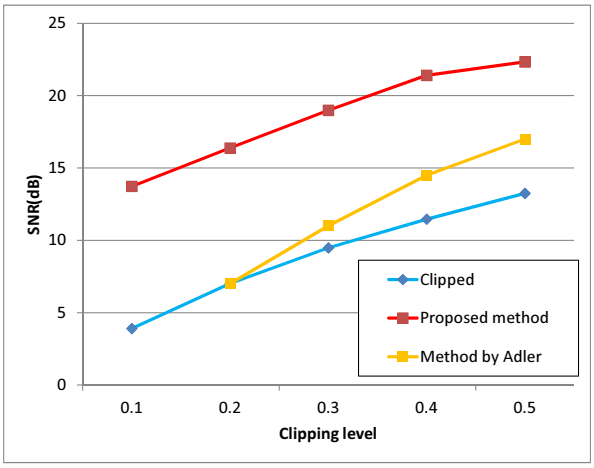

Fig. 3 SNR performance of the proposed method (music).

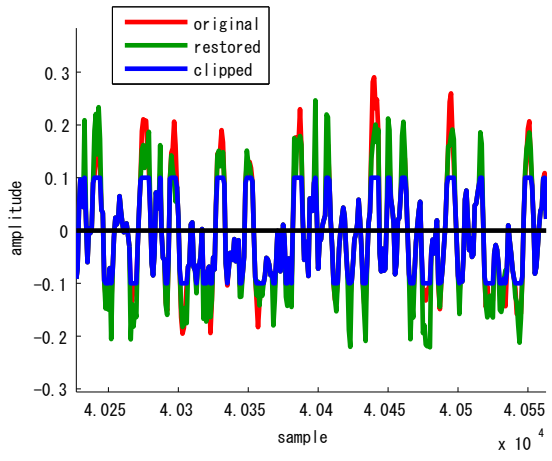

Fig. 4 An example of waveform of restored signal.

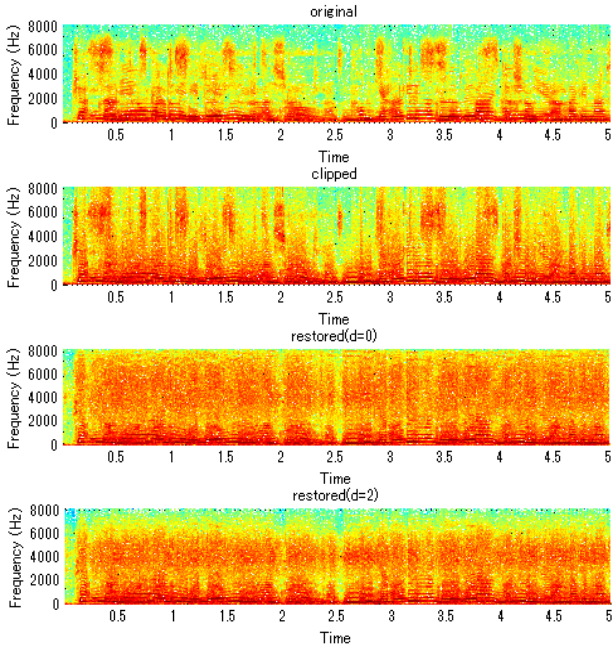

Fig. 5 An example of spectrogram of restored signal.

IEEE Trans. Acoust., Speech and Signal Process., vol. 34, no. 2, pp. 317–330, 1986.

[3] A. Dahimene, M. Noureddine, and A. Azrar, “A simple algorithm for the restoration of clipped speech signal,” Informatica, vol. 32, pp. 183–188, 2008.

[4] S. J. Godsill, P. J. Wolfe, and W. N. W. Fong, “Statistical model-based approaches to audio restoration and analysis,” J. of New Music Research, vol. 30, no. 4, 2001.

[5] 三浦晋, 中島弘史, 牧野昭二, 山田武志, 中臺一博, “クリップした音響信号の修復,” 日本音響学会春季研究発表会, pp. 941-944, Mar. 2011.

[6] S. Miura, H. Nakajima, S. Miyabe, S. Makino, T. Yamada, K. Nakadai, “Restoration of clipped audio signal using recursive vector projection,” Proc. TENCON, pp. 787–790, 2011.

[7] Hirofumi Nakajima, Mikio Tohyama, and Masashi Tanaka, ”A recurcive algorithm for digital filters to reduce the number of multipliers,” Proc. of inter・noise 96, pp. 2801-2804, 1996

[8] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, “A constrained matching pursuit approach to audio declipping,” Proc. ICASSP, pp. 329–332,2011.

[9] ITU-T Rec. P.862, “Perceptual evaluation of speech quality (PESQ) : An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” 2001.