LEAST SQUARES SIGNAL DECLIPPING FOR ROBUST SPEECH RECOGNITION

Carnegie Mellon

Mark J. Harvilla and Richard M. Stern
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA, USA

Electrical & Computer ENGINEERING

Abstract

This paper introduces a novel declipping algorithm based on constrained least-squares minimization. Digital speech signals are often sampled at 16 kHz and classic declipping algorithms fail to reconstruct accurately the signal at this sampling rate due to the scarcity of reliable samples after clipping. The Constrained Blind Amplitude Reconstruction algorithm (CBAR) interpolates missing data points such that the resulting function is smooth while ensuring the inferred data fall in a legitimate range. The inclusion of explicit constraints helps to guide an accurate interpolation. Evaluation of declipping performance is based on automatic speech recognition word error rate and Constrained Blind Amplitude Reconstruction is shown to outperform the current state-of-the-art declipping technology under a variety of conditions. Declipping performance in additive noise is also considered.

Audio Clipping

$$x_c[n] = \begin{cases} x[n] & \text{if } |x[n]| < \tau \\ \tau \cdot \operatorname{sgn} x[n] & \text{if } |x[n]| \geq \tau \end{cases}$$

Audio clipping typically occurs in one of three ways:

1. Upon recording, as a result of exceeding the dynamic range limitations of the A/D converter.
2. As a result of writing improperly-amplitude-normalized data to a file.
3. On purpose, to achieve a desirable perceptual characteristic.

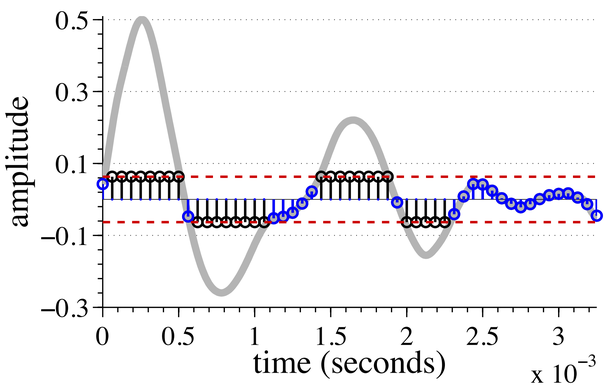


Figure 1: 16-kHz speech signal before and after clipping. The reliable samples after clipping are shown in blue,and the clipped samples are shown in black. The original unclipped signal is in gray.

Sparsity-based declipping (Kitic *et al.*)

- Using the iterative hard thresholding (IHT) algorithm, the Kitic-IHT algorithm learns a sparse representation of incoming clipped speech in terms of Gabor basis vectors.
- The learned sparse representation is then used to reconstruct clipped speech on a frame-by-frame basis.

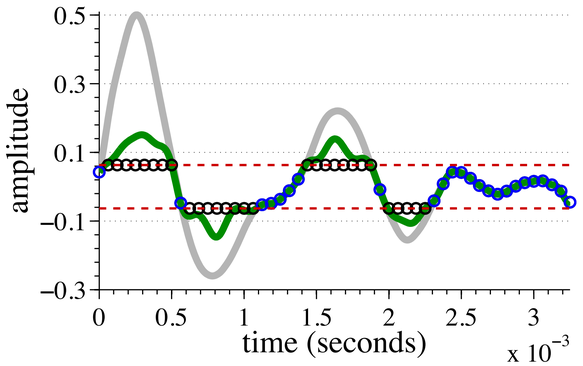


Figure 2: Illustration of the Kitic-IHT reconstruction (green).

Least squares declipping (Selesnick)

- Selesnick-LS interpolates clipped signal segments by minimizing the third derivative in the least squares sense.

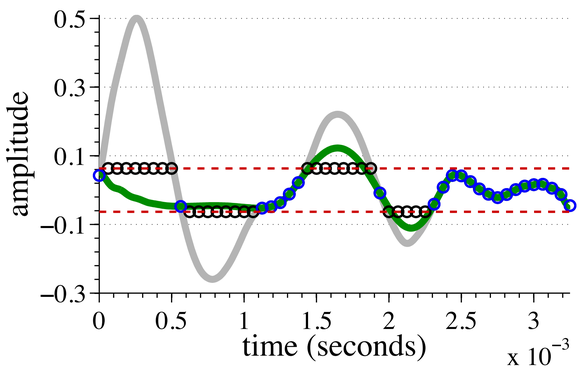


Figure 3: Illustration of the Selesnick-LS reconstruction (green); note the illegitimacy of the interpolation, i.e., it falls below τ in magnitude.

Constrained Blind Amplitude Reconstruction (CBAR)

- CBAR expands on the Selesnick technique; each speech frame is declipped by solving the following nonlinear constrained optimization problem:

$$\underset{\boldsymbol{x}_c}{\text{minimize}} \quad \left|\left|\boldsymbol{D}_2(\boldsymbol{S}_r^T\boldsymbol{x}_r + \boldsymbol{S}_c^T\boldsymbol{x}_c)\right|\right|_2^2$$

$$\text{subject to} \quad \boldsymbol{x}_c \circ \operatorname{sgn} \boldsymbol{S}_c\boldsymbol{x} \geq +\tau\mathbf{1}$$

- In the above, $\boldsymbol{D}_2$ is the 2nd-derivative operator, $\boldsymbol{S}_r$ and $\boldsymbol{S}_c$ are masking matrices that separate unclipped and clipped samples, respectively; $\boldsymbol{x}$ contains all samples of the speech frame, $\boldsymbol{x}_r$ and $\boldsymbol{x}_c$ contain unclipped and clipped samples, respectively; τ is the clipping threshold, and ∘ represents the element-wise product.
- The constrained minimization ensures that the restored samples are greater than the clipping threshold, τ, in magnitude.

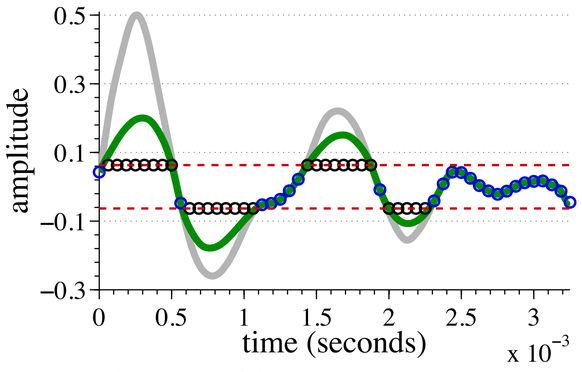
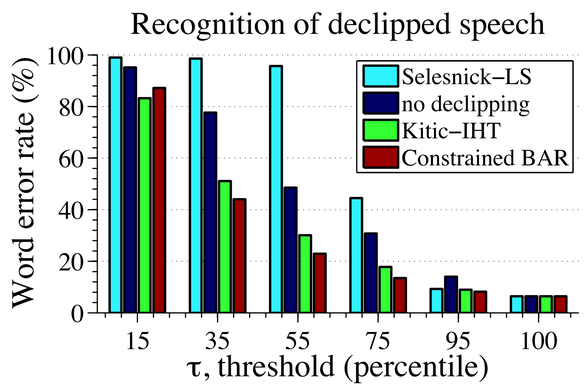


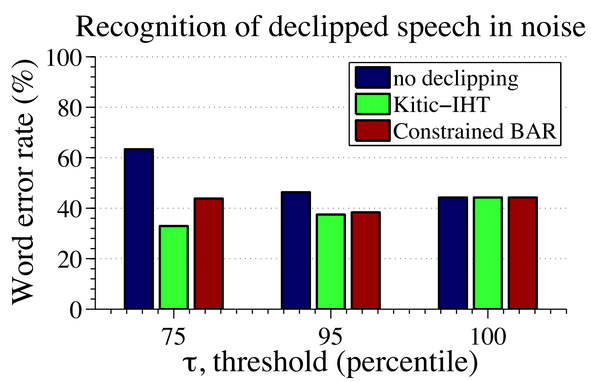
Figure 4: Illustration of the CBAR reconstruction (green); note the smoothness of the interpolation.

- Because of the inclusion of the hard constraint in the objective function, the CBAR algorithm is relatively computationally inefficient (significantly slower than Selesnick-LS and slightly slower than Kitic-IHT).
- Future research involves employing a "soft" constraint to simplify the minimization.

Experimental Results



(a) No additive noise: CBAR performs best in 80% of cases.



(b) Performance in 15-dB additive white Gaussian noise (SNR calculated after clipping).

Figure 5: Word error rates of the CMU Sphinx-III automatic speech recognition system using the DARPA RM1 database.

Summary

- Imposing a hard constraint on the minimization significantly improves reconstruction quality.
- CBAR yields the lowest WER in 80% of tested cases (all but the lowest τ).
- The presence of additive noise vitiates the effectiveness of CBAR.
- The use of a soft constraint will dramatically decrease computational complexity.