Audio restoration using plug-and-play approach

Michal Švento
dept. name of organization (of Aff.)
Brno University of Technology
Brno, Czech Republic
212584@vut.cz

Ondřej Mokrý
Signal Processing Laboratory
Brno University of Technology
Brno, Czech Republic
xmokry12@vut.cz

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—speech enhancement, deep learning, Douglas-Rachford algorithm**

I. Introduction

Audio enhancement tasks mostly face problems like missing or damaged samples, noise, or clipping. If we consider speech signal, we should not avoid the intelligibility problems. Each problem has developed its own way of enhancing the signal. Nowadays, the bestway to differentiate algorithms is with two categories: conventional (autoregressive models, sparsity-based) and solutions using deep learning.

In conventional methods dominates Janssen [1] and Etter [2]. These approaches are based on autoregressive signal modeling [3]. Sparse signal representation has changed efficiency of restoration, mainly because increase of computing power. The information hidden in frequency representation (using proper time-frequency analysis) is sparse, i.e. we do not need each spectral coefficient to repair the signal with improved subjective results. The most advanced works using sparsity are [4]–[7].

Deep learning algorithms have also made their own progress in this area. The most efficient neural network models are autoencoders, recurrent neural networks (RNNs) and Generative Adversial Network (GAN). Current state-of-the-art deep learned algorithms are Speech Enhancement GAN (SEGAN) [8], NSNet [9], FullSubNet [10].

In [11] was introduced Plug-and-Play method for image restoration. The idea of a hybrid model, combining conventional approach (convex minimization) with deep learning, has shown succesful. Our motivation is to transform this model to audio problems with minor differences. We replace Alternating Direction Multiplier Method (ADMM) with Douglas-Rachford algorithm (DR algorithm). Denoiser will be chosen from state-of-the-art audio denoisers.

This paper is organized as follows. In section II we introduce our task in mathematical view and compose minimazation task. Section III presents Plug-and-Play method and its challenges. Section IV discusses about results and further improvements of algorithm.

II. Prerequsities

In this section, we formulate the first task – inpainting. The proposed method [11] assumes any damage, but we start with missing samples and then expand the model for various damages. The rest of the section explains minimization problem solved by DR algorithm. Solution has two approaches. First, using frequency coefficients as input explained in II-B [3]. Second, using samples in time domain is described in subsection II-C [12].

A. Task formulation

We consider column vector $\mathbf{s} \in \mathbb{R}^N$ as our observed damaged single-channel signal of length N. We have set I of sample indices $\{1, 2, \ldots, N\}$, which have two subsets: I^M for missing positions and I^R stands for reliable positions. Therefore, samples $\mathbf{s}(I^M)$ are considered reliable (undamaged) and $\mathbf{s}(I^R)$ are samples, which we are looking for. It is common to rewrite it in matrix form:

$$\mathbf{s}_{\mathrm{R}} = \mathbf{M}_{\mathrm{R}}\mathbf{s},$$

where M_{R} is mask matrix $\mathbf{M}_R \in \mathbb{R}^{|I^R|\times N}$ ($|I^R|$ as length of subset), selecting rows frow indentity matrix $\mathbf{M}$ [4].

B. COEFFS INPUT Aprroach – WORK NAME

We define our task as sparsity-based problem, minimizing ℓ_0 norm of signal $\mathbf{s}$. However, this task leads to an NP-hard problem and is hardly solvable [3]. The closest redefinition is to use the ℓ_1 norm as follows:

$$\arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 \quad \text{s.t. } D\mathbf{c} \in \Gamma, \tag{1}$$

where D is synthesis operator (inverse discrete Gabor transform) and we assume A, $D = A^*$ and therefore reconstructed signal corresponds to $\mathbf{y} \approx D\mathbf{c}$. Set Γ is defined as follows:

$$\Gamma = \{\mathbf{y} \in \mathbb{R}^L \mid M_{\mathrm{R}}\mathbf{y} = M_{\mathrm{R}}\mathbf{s}\}, \tag{2}$$

One of the suitable solutions is DR algorithm in 1 [3].

Operator $\mathrm{proj}_\Gamma(arg)$ is projection onto convex set Γ and $\mathrm{soft}_\gamma(arg)$ is soft thresholding operator. Both are proximal operators.

The condition $\delta \in (0, 1)$ is strict for convergence of solution [13].

Identify applicable funding agency here. If none, delete this.

Algorithm 1 Douglas-Rachford algorithm – model with frequency coefficients

```
Input: $\gamma > 0$, $\delta \in (0,1)$ ,
for $n = 0, 1, \dots$ do
    $\widetilde{\mathbf{c}}_n = \mathrm{proj}_\Gamma(\mathbf{c}_n)$
    $\mathbf{c}_{n+1} = \mathbf{c}_n + \lambda\left(\mathrm{soft}_\gamma\left(2\widetilde{\mathbf{c}}_n - \mathbf{c}_n\right) - \widetilde{\mathbf{c}}_n\right)$
end for
return $D(\mathrm{proj}_\Gamma(\mathbf{c}_n))$
```

C. SAMPLES as INPUT APPROACH

Second approach uses time domain samples as input. It has mainly computational advice against first approach, in algorithm are less time-frequency transformations (e.g less multiplication operations per iteration). Problem is, that we do not know proximal operator of ℓ_1 norm after analysis, but we can use approximal operator [12].

The main minimazation task reformulates as:

$$\arg\min_{\mathbf{s}} \|\mathbf{s}\|_1 \quad \text{s.t. } \mathbf{s} \in \Gamma, \tag{3}$$

and Γ is similarly as in (eqref2)

$$\Gamma = \{\mathbf{s} \in \mathbb{R}^N \mid M_\mathrm{R}\mathbf{y} = M_\mathrm{R}\mathbf{s}\}. \tag{4}$$

Afterwards our algorithm resolves to following:

Algorithm 2 Douglas-Rachford algorithm – model with time coefficients

```
Input: $\lambda > 0$, $\gamma > 0$, $\mathbf{x}_0 \in \mathbb{R}^N$
for $n = 0, 1, \dots$ do
    $\widetilde{\mathbf{x}}_n = \mathrm{proj}_\Gamma(\mathbf{x}_n)$
    $\mathbf{x}_{n+1} = \mathbf{x}_n + \lambda\left(D\left(\mathrm{soft}_\gamma\left(A\left(2\widetilde{\mathbf{x}}_n - \mathbf{x}_n\right)\right)\right) - \widetilde{\mathbf{x}}_n\right)$
end for
return $\mathrm{proj}_\Gamma(\mathbf{x}_n)$
```

Projection, in this case, is simply replacing reconstructed samples in positions considered reliable.

III. Plug-and-Play inpainting

A. general algorithm

Algorithm 3 Plug-and-Play DR algorithm

```
Input: in
for $n = 0, 1, \dots$ do
    $\widetilde{\mathbf{x}}_n = \mathrm{proj}_\Gamma(\mathbf{x}_n)$
    $\mathbf{x}_{n+1} = \mathbf{x}_n + \lambda\left(\mathcal{D}\left(2\widetilde{\mathbf{x}}_n - \mathbf{x}_n\right) - \widetilde{\mathbf{x}}_n\right)$
end for
return $\mathrm{proj}_\Gamma(\mathbf{x}_n)$
```

B. choice of denoiser

C. Denoisers

IV. Testing data and evaluation

V. Conclusion

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

[1] A. Janssen, R. Veldhuis, and L. Vries, “Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 317–330, Apr. 1986.

[2] W. Etter, “Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters,” *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1124–1135, May 1996.

[3] O. Mokrý and P. Rajmic, “Audio Inpainting: Revisited and Reweighted,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2906–2918, 2020.

[4] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley, “Audio Inpainting,” *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, pp. 922–932, Mar. 2012.

[5] S. Kitić, N. Bertin, and R. Gribonval, “Sparsity and cosparsity for audio declipping: a flexible non-convex approach,” Tech. Rep., Jun. 2015, arXiv:1506.01830 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1506.01830

[6] P. Záviška, P. Rajmic, O. Mokrý, and Z. Průša, “A Proper Version of Synthesis-based Sparse Audio Declipper,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 591–595, iSSN: 2379-190X.

[7] O. Mokrý, P. Záviška, P. Rajmic, and V. Veselý, “Introducing SPAIN (SParse Audio INpainter),” in *2019 27th European Signal Processing Conference (EUSIPCO)*, Sep. 2019, pp. 1–5, iSSN: 2076-1465.

[8] S. Pascual, A. Bonafonte, and J. Serrà, “SEGAN: Speech Enhancement Generative Adversarial Network,” Tech. Rep., Jun. 2017, arXiv:1703.09452 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1703.09452

[9] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, “Weighted Speech Distortion Losses for Neural-network-based Real-time Speech Enhancement,” Tech. Rep., Feb. 2020, arXiv:2001.10601 [cs, eess] type: article. [Online]. Available: http://arxiv.org/abs/2001.10601

[10] X. Hao, X. Su, R. Horaud, and X. Li, “FullSubNet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement,” Tech. Rep., Jan. 2021, arXiv:2010.15508 [cs, eess] type: article. [Online]. Available: http://arxiv.org/abs/2010.15508

[11] S. H. Chan, X. Wang, and O. A. Elgendy, “Plug-and-Play ADMM for Image Restoration: Fixed Point Convergence and Applications,” Tech. Rep., Nov. 2016, arXiv:1605.01710 [cs] type: article. [Online]. Available: http://arxiv.org/abs/1605.01710

[12] O. Mokrý and P. Rajmic, “Approximal operator with application to audio inpainting,” *Signal Processing*, vol. 179, p. 107807, Feb. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168420303510

[13] P. L. Combettes and J.-C. Pesquet, “Proximal Splitting Methods in Signal Processing,” Tech. Rep., May 2010, arXiv:0912.3522 [math] type: article. [Online]. Available: http://arxiv.org/abs/0912.3522