Joint audio denoising and inpainting with plug-and-play proximal algorithm

Michal Švento
Brno University of Technology, FEEC,
Department of Telecommunications,
Technická 12, 616 00 Brno, Czech Republic
212584@vut.cz

Ondřej Mokrý
Brno University of Technology, FEEC,
Department of Telecommunications,
Technická 12, 616 00 Brno, Czech Republic
xmokry12@vut.cz

***Abstract*—We propose plug-and-play variant of the Douglas–Rachford proximal algorithm for audio inpainting, which replaces a proximal step with a denoiser. In our situation, the observed samples are further degraded by noise. We demonstrate that the plug-and-play aproach has potential to succeed in this joint task of inpainting and denoising. Objective metrics show that the new method outperforms a conventional counterpart and that it is less sensitive to model hyperparameters.**

***Index Terms*—speech enhancement, deep learning, denoising, Douglas–Rachford algorithm, inpainting**

I. Introduction

Audio enhancement tasks mostly face problems like missing or damaged samples, noise, or clipping. Considering speech signals, we are not only interested in restoring the degradation sample by sample, but we also aim at improving the intelligibility of the recorded speech. Each restoration problem has developed its own way of enhancing the signal. Nowadays, the best way to differentiate algorithms is into two categories: conventional, e.g. using autoregressive (AR) modeling or sparsity-based optimization, and solutions using deep learning. The present paper focuses on the case of restoring a partially observed signal whose observed samples are further degraded by noise, i.e., the aim is to perform simultaneous inpainting and denoising of the speech signal.

In conventional approaches to inpainting, the AR-based Janssen [1] and Etter [2] algorithms dominate in terms of restoration quality. A more recent, successful class of methods is based on sparsity. The key idea is that after performing proper time-frequency analysis of an audio signal, most of the information is concentrated in a few coefficients, i.e., it is sparse. This can be applied as fitting the sparsest possible restoration either to the reliable observed samples [3]–[6], or to a signal not much diverging from the observation in the case of denoising [7].

Deep learning algorithms have also made their own progress in this area. The most efficient neural network models are autoencoders, recurrent neural networks (RNN) and Generative Adversial Networks (GAN). Current state-of-the-art deep learned algorithms are base on Speech Enhancement GAN (SEGAN) [8], NSNet [9], FullSubNet [10]. While learning-based algorithms allow to adapt to real-world signals, rather than to rely on hand-crafted priors like sparsity or the AR nature of signals, they need large datasets for training. Furthermore, neural networks are usually trained for a specific problem, lacking universal applicability on similar restoration tasks, in contrast to sparsity-based methods [11]–[13].

As a compromise between the conventional and learning-based methods, the plug-and-play method for image restoration was introduced in [14] and subsequently studied in [15], where part of each iteration of an optimization algorithm is replaced by a (learned) denoiser. In the present paper, we propose a hybrid algorithm based on the same paradigm, aiming at restoration of degraded speech. While [15] focused on adapting the Alternating Direction method of Multipliers (ADMM) and a recent declipping approach used the learned element only partially [16], we choose an opposite approach by working with a simple Douglas–Rachford algorithm (DRA) and exploring the trade-off between data fitting and denoising in the algorithm.

The paper is organized as follows. In section II we introduce the task from mathematical point of view and we define the restoration as a minimization task. Section III presents the plug-and-play method and its challenges. Section IV discusses the results and further improvements of algorithm. Finally, Section V concludes the paper.

II. Prerequsities

In this section, we formalize the task of inpainting and denoising and propose algorithmic solutions based on DRA. First, using frequency coefficients as input explained in II-B [17]. Second, using samples in time domain is described in subsection II-C [18].

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 of sample indices $\{1, 2, \ldots, N\}$, which has two disjunctive subsets: I_{M} for missing positions and I_{R} stands for reliable positions. Usually, samples $\mathbf{s}(I_{\mathrm{R}})$ are considered reliable (undamaged) and $\mathbf{s}(I_{\mathrm{M}})$ 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 $\mathbf{M}_\mathrm{R} \in \mathbb{R}^{|I_\mathrm{R}| \times N}$ is the mask matrix, selecting rows from identity matrix corresponding to the indices in I_R [3], with $|I_\mathrm{R}|$ denoting the number of indices in the subset I_R. In words, $\mathbf{M}_\mathrm{R}\mathbf{s}$ represents choosing the samples from $\mathbf{s}$ on positions I_R.

It is convenient to define the set of signals fitting the observation as

$$\Gamma = \{\mathbf{x} \in \mathbb{R}^L \mid \mathbf{M}_\mathrm{R}\mathbf{x} = \mathbf{M}_\mathrm{R}\mathbf{s}\}. \tag{1}$$

In the noise-less case, we would search for a suitable signal in Γ. On the other hand, when the observed samples are distorted by noise, we only require the solution $\mathbf{x}$ to be close to Γ.

B. Synthesis-based formulation

We define the task of finding a suitable signal in (or close to) Γ as a sparsity-based problem, minimizing ℓ_0-norm of the Gabor coefficients, i.e., the number of non-zero coefficients of the time-frequency representation of the signal. However, this task leads to an NP-hard problem and is hardly solvable [17]. The closest redefinition is to use the ℓ_1-norm as follows:

$$\arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 \quad \text{subject to} \quad \mathbf{D}\mathbf{c} \in \Gamma, \tag{2}$$

where $\mathbf{D}$ is synthesis operator (inverse discrete Gabor transform), which is the adjoint of the analysis operator $\mathbf{A}$, i.e., $\mathbf{D} = \mathbf{A}^*$. Therefore the restored signal corresponds to $\mathbf{x} = \mathbf{D}\mathbf{c}$.

In the presence of noise, the condition $\mathbf{D}\mathbf{c} \in \Gamma$ is not beneficial, since it forces the solution to still contain the noise. Its reduction can be included in the formulation by rather minimizing the distance of the solution from the set Γ. This can be computed as the difference of the solution $\mathbf{D}\mathbf{c}$ and the observation $\mathbf{s}$ on the positions I_R. Typically, squared ℓ_2-norm is used in this context:

$$\arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 + \frac{\alpha}{2}\|\mathbf{M}_\mathrm{R}\mathbf{D}\mathbf{c} - \mathbf{M}_\mathrm{R}\mathbf{s}\|_2^2. \tag{3}$$

The parameter $\alpha > 0$ manages the trade-off between sparsity of the coefficients and the data-fidelity. Both (2) and (3) could be solved using DRA [13], [17]. However, the algorithms use the Gabor coefficients as the main variable, which makes the incorporation of a denoiser (working with a signal as the input) problematic.

C. Analysis-based formulation

Second approach uses time domain samples as input and also as variables of the optimization task. In case the transform is redundant, i.e., we have more coefficients than signal samples, this second approach leads to different solutions in general [17].

The main minimazation task (4) is reformulated as,

$$\arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{x} \in \Gamma, \tag{4}$$

or, in the denoising case,

$$\arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x}\|_1 + \frac{\alpha}{2}\|\mathbf{M}_\mathrm{R}\mathbf{x} - \mathbf{M}_\mathrm{R}\mathbf{s}\|_2^2. \tag{5}$$

If we want to employ DRA also on the analysis-based task, a problem occurs that we do not know proximal operator of ℓ_1-norm after analysis, which is essential for the DRA. However, we can resort to the so-called approximal operator without significant loss of restoration quality [18]. The algorithm is summarized in Alg. 1.

Algorithm 1 DRA for (4) or (5).

Input: $\lambda_n > 0$, $\gamma > 0$, $\widetilde{\mathbf{x}}_0 \in \mathbb{R}^N$, $\beta \in [0, 1]$
1: **for** $n = 0, 1, \ldots$ **do**
2: $\quad \mathbf{x}_n = (1-\beta)\widetilde{\mathbf{x}}_n + \beta\,\mathrm{proj}_\Gamma(\widetilde{\mathbf{x}}_n)$
3: $\quad \widetilde{\mathbf{x}}_{n+1} = \mathbf{x}_n + \lambda_n\left(\mathbf{D}\left(\mathrm{soft}_\gamma\left(\mathbf{A}\left(2\mathbf{x}_n - \widetilde{\mathbf{x}}_n\right)\right)\right) - \mathbf{x}_n\right)$
4: **end for**
5: **return** $\mathbf{x}_n$

The operator proj_Γ is projection onto convex set Γ and soft_γ is soft thresholding operator [19]. Projection, in this case, means replacing restored samples in positions considered reliable by the observed samples. The parameter β allows to differentiate between simple inpainting (4) and the joint problem (5). The case of $\beta = 1$ corresponds to performing projection in the update, in line with [17]. For $\beta < 1$, it can be derived (see e.g. [19, Sec. 4 and Tab. 1]) that the update corresponds to the proximal operator of $f(\mathbf{x}) = \frac{\alpha}{2}\|\mathbf{M}_\mathrm{R}\mathbf{x} - \mathbf{M}_\mathrm{R}\mathbf{s}\|_2^2$ with $\alpha = \frac{\beta}{1-\beta}$.

III. Plug-and-play inpainting

The plug-and-play method was proposed in [14] and [15] for image restoration. The main idea is to replace a proximal operator inside an iterative algorithm with an denoiser. Since this approach does not affect the data-fitting part of the minimization, it should be suitable for different resotaration tasks. We rewrite this task to solve our audio inpainting problem using DRA with approximal operator (analysis-based), i.e., we modify Alg. 1, leading to Alg. 2:

Algorithm 2 Plug-and-play DRA

Input: $\lambda_n > 0$, $\gamma > 0$, $\widetilde{\mathbf{x}}_0 \in \mathbb{R}^N$, $\beta \in [0, 1]$
1: **for** $n = 0, 1, \ldots$ **do**
2: $\quad \mathbf{x}_n = (1-\beta)\widetilde{\mathbf{x}}_n + \beta\,\mathrm{proj}_\Gamma(\widetilde{\mathbf{x}}_n)$
3: $\quad \widetilde{\mathbf{x}}_{n+1} = \widetilde{\mathbf{x}}_n + \lambda_n\left(\mathcal{D}\left(2\mathbf{x}_n - \widetilde{\mathbf{x}}_n\right) - \mathbf{x}_n\right)$
4: **end for**
5: **return** $\mathbf{x}_n$

Convergence of the plug-and-play approach can be proven in case $\mathcal{D}$ is a non-expansive operator [15]. However, this is hard to prove in practice with off-the-shelf denoisers. Thus, the only requirement for our denoiser $\mathcal{D}$ was primarily good results in subjective metrics (discussed more in IV-A). The implementation of the denoiser of our choice is based on the *mayavoz* toolkit [20]. This tool provides us simple use of learned model with downloadable checkpoints. Our selection is the WaveUnet learned on the Valentini dateset [21] (the variant with 28 speakers).

IV. Testing data and evaluation

As introduced before, we use discrete Gabor transform as the analysis operator $\mathbf{A}$ with following setup parameters:

Gauss window with length $w = 1024$, hop length $a = 256$ and number of frequency channels used in Fast Fourier transform $n_{\text{FFT}} = 1024$. To satisfy the Parseval tight frame condition, the synthesis operator $\mathbf{D} = \mathbf{A}^*$ is configured the same way [17].

A. Metrics

A standard metric for restoration quality is the signal-to-noise ratio (SNR), where the *signal* represents the clean ground truth and *noise* is the difference of the restoration and the ground truth. We calculate SNR either for whole signal or only on I_{M}, i.e, on the indices of missing samples. The higher the SNR, the better the restoration.

A drawback of SNR is that it measures sample-wise difference, which does not take into account human perception of sound quality. Valuable metrics for evaluating speech signals are those simulate measure human perception. We use two objective metrics: Perceptual Evaluation of Speech Quality (PESQ) [22] and Short-Time Objective Intelligibility (STOI) [23]. PESQ has scale from -0.5 to 4.5, with higher score meaning better result. STOI measures correlation of two signals on a scale from 0 to 1.

B. Dataset and degradation setup

We have chosen to work with the Valentini dataset [21] used commonly in speech enhancement challenges (the dataset is split to two subsets: clean and noisy). For each signal, we used the clean sample as a reference and its noisy variant as the input to the algorithm. The average SNR of our files was approximately 11 dB. As a simulation of inpainting task, we used random drop-out of 40% of samples as used in [18]. This means that the reliable set I_{R} consisted of 60% of all the sample indices.

C. Choice of the soft threshold

To reduce the number of variables in the comparison between the conventional and plug-and-play approach, we decided to fix the γ parameter in DRA. We examined DRA with values of $\gamma \in \{0.001, 0.01, 0.1\}$ for a single restoration task. At the same time, we examined the effect of the trade-off parameter β in Alg. 1, since it has a significant effect on the results in the case of signal denoising. Tested signals are pair of clean and noisy signal from the dataset described above with random drop-out of 40% samples. The parameters of DRA were: $\lambda_{n+1} = 0.9 \cdot \lambda_n$, starting with $\lambda_0 = 1$, and 50 iterations. The results are shown in Fig. 1:

The best results were obtained with $\gamma = 0.01$, which reached the best restoration quality, especially with increasing ratio of projected signal.

D. Test setup

We have chosen randomly 10 audio files for final evaluation from the Valentini dataset [21] described before. As justified in IV-C, we have chosen $\gamma = 0.01$. The simulation was split to three restoration variants: two conventional ones and one including the learned denoiser. The conventional ones were

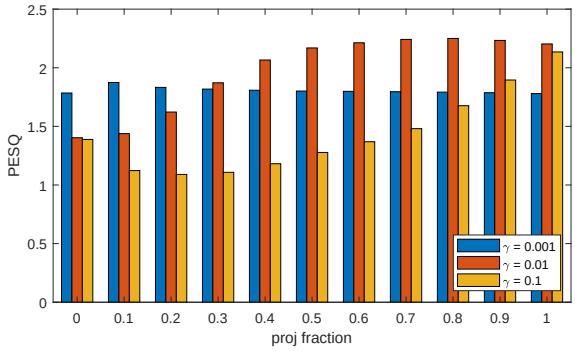

Fig. 1: PESQ metric results made to choose the best value of γ in DRA for testing with multiple audio files (these values are shown in the legend). The results in terms of STOI and SNR were principally the same. The label *proj fraction* refers to parameter β from Alg. 1.

two versions of DRA 1 with same parameters except for the number of iterations (50 or 500) and the choice of λ_n (decreasing or constant). The DRA with denoiser proposed in Alg. 2 was tested with 50 iterations, as in the shorter conventional method. In all the algorithms, the initial λ_0 was set to 1. For the case of DRA 1 with 500 iterations, we kept constant $\lambda_n = 0.1$, which provided better results compared to the choice of $\lambda_n = 1$. In the case of the plug-and-play version 2, and the reference DRA 1 with 50 iterations, the parameter was set to decrease in every iteration as $\lambda_{n+1} = 0.9 \cdot \lambda_n$. The motivation for decreasing λ is clear from Fig. 2. The test was performed using DRA and every option should eventually converge to the same optimal point in time [19]. However, the choice has an impact in practice, either in particular choice of the argument of the minima, or due to numerical reasons and finite number of iterations. If λ decreases, we minimize the impact of soft threshold operator in latter iterations which appears to be effective in our case.

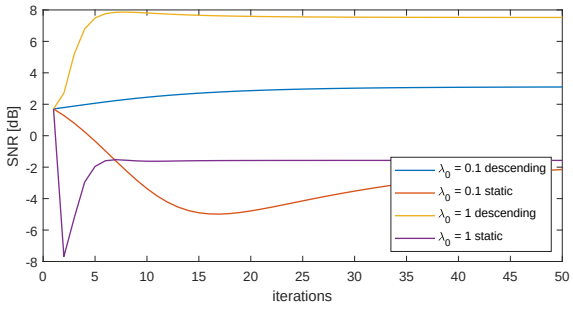

Fig. 2: SNR from DRA tested for 50 iterations with static and decreasing choice of $\lambda \in \{0.1, 1\}$. Tested signals are same as in experiment with γ.

E. Comparison results

Graphs with STOI, PESQ and SNR are presented in Fig. 3. We see that plug-and-play is less sensitive to the fraction

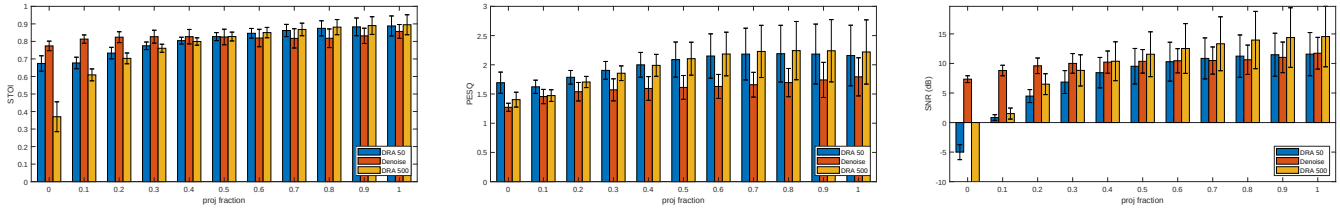

Fig. 3: Average results from 10 tested signals. Metrics are from left to right: STOI, PESQ, SNR.

of projection compared to DRA, and in comparison with DRA (50 iterations, decreasing λ_n), it reaches better result. DRA with 500 iterations has higher variance of the results, but globally, it outperformed the plug-and-play method with optimal choice of the parameters such as the ratio β.

V. Conclusion

Our reformulation of plug-and-play method for audio inpainting with noisy data was demonstrated to be successful with properly chosen ratio of projection and λ. Next steps could be to extend the test for other similar tasks (declipping, dequantization), following the pattern of root method [15], solving multiple tasks with one setup of the algorithm. We also suggest suggested not to use off-the-shelf denoiser, but learn own model on data related to our problem (e.g. specific noise due to dropout of a certain percentage of samples). In addition, we might be able to make the self-learned denoiser non-expansive, thus satisfying the theoretical convergence guarantees [14], [15]. A compromise might be to use transfer learning, i.e., to take a successful denoiser and continue learning from a checkpoint with our dataset, with idea to reduce artifacts generated in the inpainting task.

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, 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, pp. 1124–1135, May 1996.

[3] 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.

[4] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: a flexible non-convex approach," in *LVA/ICA 2015 – The 12th International Conference on Latent Variable Analysis and Signal Separation*, (Liberec, Czech Republic), pp. 243–250, Aug. 2015.

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

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

[7] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social sparsity! neighborhood systems enrich structured shrinkage operators," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2498–2511, 2013.

[8] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," tech. rep., June 2017. arXiv:1703.09452 [cs] type: article.

[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.

[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.

[11] C. Gaultier, N. Bertin, S. Kitić, and R. Gribonval, "A modeling and algorithmic framework for (non)social (co)sparse audio restoration," 2017.

[12] O. Mokrý, P. Rajmic, and P. Záviška, "Flexible framework for audio reconstruction," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx2020)*, vol. 1, (Vienna, Austria), Sept. 2020-21.

[13] P. Záviška, P. Rajmic, and O. Mokrý, "Audio dequantization using (co)sparse (non)convex methods," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toronto, Canada), pp. 701–705, 2021.

[14] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *2013 IEEE Global Conference on Signal and Information Processing*, IEEE, Dec. 2013.

[15] 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.

[16] T. Tanaka, K. Yatabe, M. Yasuda, and Y. Oikawa, "APPLADE: Adjustable plug-and-play audio declipper combining DNN with sparse optimization," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, may 2022.

[17] 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.

[18] O. Mokrý and P. Rajmic, "Approximal operator with application to audio inpainting," *Signal Processing*, vol. 179, p. 107807, Feb. 2021.

[19] P. Combettes and J. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, vol. 49, pp. 185–212, 2011.

[20] A. Kiran, "shahules786/mayavoz," Mar. 2023. original-date: 2022-08-20T04:44:35Z.

[21] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," *http://parole.loria.fr/DEMAND/*, Aug. 2017.

[22] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752 vol.2, May 2001. ISSN: 1520-6149.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217, Mar. 2010. ISSN: 2379-190X.