Janssen 2.0: Audio Inpainting in the Time-frequency Domain

Ondřej Mokrý
Dept. of Telecommunications
Brno University of Technology
Czech Republic
ondrej.mokry@vut.cz

Peter Balušík
Dept. of Telecommunications
Brno University of Technology
Czech Republic
230531@vut.cz

Pavel Rajmic
Dept. of Telecommunications
Brno University of Technology
Czech Republic
pavel.rajmic@vut.cz

***Abstract*—The paper focuses on inpainting missing parts of an audio signal spectrogram, i.e., estimating the lacking time-frequency coefficients. The autoregression-based Janssen algorithm, a state-of-the-art for the time-domain audio inpainting, is adapted for the time-frequency setting. This novel method, termed Janssen-TF, is compared with the deep-prior neural network approach using both objective metrics and a subjective listening test, proving Janssen-TF to be superior in all the considered measures.**

***Index Terms*—audio inpainting, autoregression, deep prior, DPAI, time-frequency, spectrogram.**

I. Introduction

Audio inpainting, sometimes referred to as missing audio interpolation or concealment, is the task of filling in the absent parts of audio with a meaningful content. Typically, the missing information is in the form of lost samples in the *time domain*. If compact regions of such samples are present, they are often referred to as the gaps. For the described task, numerous methods have been developed: those exploiting the autoregressive character of audio [1], [2], methods based on the sparsity of the time-frequency representation of audio [3], [4], [5], [6], or methods relying on similarities in audio recordings [7], [8], [9]. Quite recently, deep learning audio inpainting methods appeared [10], [11]. For gaps of up to 50 ms in length, the Janssen method [2], [12] has been evaluated the best for a long time, but PHAIN [13] may become a new state-of-the art.

The time-domain inpainting represents the most natural case from the point of view of the application to signal recovery. Nonetheless, many practical algorithms, such as classic audio codecs (MP3, AAC, etc.), operate on short-time signal spectra. From this perspective, it is valuable to think about inpainting in the *time-frequency domain*,[1] besides the fact that it is an interesting problem per se. In the case under consideration, a portion of time-frequency (TF) coefficients[2] is missing, which need to be recovered to allow a later reconstruction of the time-domain signal. Indeed, there have been a few methods developed in the TF context, all of them relying on neural modeling [16], [17], [18]. The most recent work [18] is based on the concept of a "deep prior", a neural network that does not need training data besides the corrupted observation [19].

The goal of this paper is to adapt the Janssen algorithm [2], proven successful in the time-domain case, for the TF-domain case, and compare its performance with the recent deep-prior-based approach of [18], both using objective metrics and by listening tests.

The work was supported by the Czech Science Foundation (GAČR) Project No. 23-07294S. The authors are grateful to NVIDIA for their donation of the Titan XP graphic card.

[1]In other fields, one can go even further and require inpainting purely in the frequency domain, see for example [14], [15].

[2]typically in the form of a complex-valued matrix

II. Gaps and the Short-time Fourier Transform

Note that a compact gap in the time domain is far from being equivalent to a gap in the TF domain, which can be associated with the uncertainty principle. When a time-domain signal is transformed using the Short-time Fourier Transform (STFT), it is first divided into overlapping frames, which are weighted using windows such as the Hann window (see, e.g., [20]). When a gap is present, the number of TF coefficients affected by the gap depends on how much the frames overlap. In practical scenarios, the gap overlaps with more than one frame; setting the missing samples to zeros then reduces the value of the corresponding TF coefficients, leading to a kind of TF-fadeout. For wide gaps, even a number of completely zero columns can appear in the TF domain (i.e., in the spectrogram), see the analysis in [21]. Since the inverse STFT also uses windowed frames, the same holds in the opposite direction: consecutive empty spectrogram columns result in a decrease in the time-domain amplitude, governed by the window shape. In turn, there is no way how to equitably compare time-domain inpainting techniques with those specialized for the TF-domain.

III. Deep Prior Audio Inpainting (DPAI)

The deep prior, introduced in [19], is a concept used for the reconstruction of corrupted signals using a convolutional neural network (CNN). In contrast to most CNN applications, [19] does not utilize a training dataset; the reconstruction only involves a proper CNN architecture and the corrupted signal. The deep prior concept has also proven that the network architecture matters significantly more than previously thought.

A. Application of Deep Prior to Audio Inpainting

The deep prior reconstruction can be utilized in a variety of applications, as outlined in [19]. In the context of the time-domain audio inpainting, consider a corrupted signal $\mathbf{x}^{\text{cor}} \in \mathbb{R}^N$, having missing regions (filled with zeros), compared with the original signal $\mathbf{x}^{\text{orig}} \in \mathbb{R}^N$. Formally, if a binary mask $\mathbf{m} \in \{0,1\}^N$ is utilized to represent the missing sections of the signal, the relation $\mathbf{x}^{\text{cor}} = \mathbf{m} \odot \mathbf{x}^{\text{orig}}$ holds, where $\odot$ represents the elementwise product. Such a relation naturally characterizes the feasible set $\mathit{\Gamma}_{\text{T}} = \{\mathbf{x} \,|\, \mathbf{m} \odot \mathbf{x} = \mathbf{m} \odot \mathbf{x}^{\text{orig}}\}$.

An encoder–decoder CNN denoted $f_\theta(\mathbf{n})$ is assumed to be able to generate *any* signal of size N using a random but fixed input noise $\mathbf{n} \in \mathbb{R}^N$, using a proper θ. This is a reliable assumption since the number of network parameters θ (weights and biases) is significantly greater than N.

In the deep prior reconstruction, also the parameters θ are initialized randomly. To obtain a reconstruction result, θ is then optimized as follows:

$$\theta^* = \operatorname*{argmin}_\theta E(\mathbf{m} \odot f_\theta(\mathbf{n}), \mathbf{x}^{\text{cor}}), \tag{1}$$

where E is a differentiable data fidelity function measuring the proximity of its two inputs. The optimal θ^* can be obtained using an iterative optimizer, e.g., Adam [22]. The crucial observation is that the optimization must be interrupted before it reaches θ^*, for which we would have $f_{\theta^*}(\mathbf{n}) = \mathbf{x}^{\text{cor}}$, meaning basically overfitting. Instead, early stopping of the training is desired, resulting in an estimate $\hat{\theta}$ and a corresponding $\hat{\mathbf{x}} = f_{\hat{\theta}}(\mathbf{n})$ [19].

The authors of [18] used an analog approach to inpaint audio in the time-frequency domain. Through the use of the STFT, denoted $\mathbf{L}$ in the following, an audio recording $\mathbf{x}$ is considered merely in the TF domain, where it forms a spectrogram matrix $\mathbf{L}\mathbf{x} = \mathbf{X} \in \mathbb{C}^{T \times F}$. Therefore the CNN operates on spectrograms and $\hat{\mathbf{X}} = f_{\hat{\theta}}(\mathbf{N})$ for a random but fixed matrix $\mathbf{N}$. Once $\hat{\mathbf{X}}$ has been established, the time-domain solution is obtained by the inverse STFT.

Regarding the mask, in [18] only the case of completely missing columns of the spectrogram is considered, i.e., for the TF mask $\mathbf{M} \in \{0,1\}^{T \times F}$, each of its columns is either full of ones or zeros. Nevertheless, a generalization to other TF masks is straightforward. The observed spectrogram $\mathbf{X}^{\text{cor}}$ is finally modeled as $\mathbf{X}^{\text{cor}} = \mathbf{M} \odot \mathbf{X}^{\text{orig}}$, where $\mathbf{X}^{\text{orig}}$ corresponds to $\mathbf{x}^{\text{orig}}$ in the time domain, i.e., $\mathbf{X}^{\text{orig}} = \mathbf{L}\mathbf{x}^{\text{orig}}$. Note that by analogy with the time-domain case, we obtain the feasible set $\mathit{\Gamma}_{\text{TF}} = \{\mathbf{X} \,|\, \mathbf{M} \odot \mathbf{X} = \mathbf{M} \odot \mathbf{X}^{\text{orig}}\}$.

In [18], the loss function E in (1) linearly combines the classic Mean Squared Error between spectrograms and the multi-scale spectrogram loss [23], [24].

B. Architecture

The authors of [18] used an architecture similar to U-Net called MultiResUNet [25]. It is a common CNN architecture with several layers; the skip connections between the encoder and decoder stages play a significant role in audio inpainting. The architecture additionally includes the so-called harmonic convolution [26], which proved to be the best among multiple tested architectures in [18]. The harmonic convolution is a variation on the common 2D convolution that takes into account higher harmonics in audio.

C. Variants

The DPAI intrinsically reconstructs the whole $\hat{\mathbf{X}}$. Yet, considering that only a part of the spectrogram was not observed, the authors of [18] evaluate the performance of their method after injecting the reliable coefficients back to the solution as a postprocessing step. Formally, such an action is a projection

$$\operatorname{proj}_{\mathit{\Gamma}_{\text{TF}}}(\hat{\mathbf{X}}) = \mathbf{M} \odot \mathbf{X}^{\text{cor}} + (\mathbf{1} - \mathbf{M}) \odot \hat{\mathbf{X}} \tag{2}$$

onto the set $\mathit{\Gamma}_{\text{TF}}$. Note that a resulting sharp transition between the regions is smeared in the time domain due to the nature of the STFT [21]; see also Sec. II. We call the reconstructions modified by (2) *with context*, while the others using the entire $\hat{\mathbf{X}}$ are *without context*. In the experimental part, we will study whether employing the context brings an improvement in performance.

IV. Janssen spectrogram inpainting (Janssen-TF)

A signal $\mathbf{x} = [x_1, \ldots, x_N]^\top \in \mathbb{R}^N$ corresponds to an autoregressive process if the following holds:

$$\sum_{i=1}^{p+1} a_i x_{n+1-i} = e_n, \quad n = 1, \ldots, N+p. \tag{3}$$

Here, $\mathbf{a} = [1, a_2, \ldots, a_{p+1}]^\top$ is the vector of model parameters (AR coefficients), p is the model order, and $\mathbf{x}$ is zero-padded such that its final length is $N + p$. The error term $\mathbf{e} \in \mathbb{R}^{N+p}$ represents a realization of a zero-mean white noise process with variance σ^2 [27, Def. 3.1.2].

In the *time-domain* inpainting method proposed by Janssen et al. [2], the restored audio signal is supposed to follow the AR model (3) while both the missing samples in the time-domain and the unknown AR parameters $\mathbf{a}$ are being estimated. Janssen et al. approach the problem iteratively by performing two principal steps in the i-th iteration:

1) for a temporary solution $\mathbf{x}^{(i)}$, estimate the AR model parameters $\mathbf{a}^{(i)}$;
2) given the model parameters $\mathbf{a}^{(i)}$, estimate a new temporary solution $\mathbf{x}^{(i+1)}$ which fits the observed samples and the norm of the model error $\mathbf{e}$ is minimized.

Step 1 can be solved efficiently using standard methods, such as the Durbin–Levinson recursion (see for example the `lpc` function in Matlab). Step 2 boils down to a mean squares problem which can be treated e.g. via matrix inversion.

The proposed adaptation of the Janssen iteration to the TF setting, denoted Janssen-TF, lies in modifying step 2. In our case, the *signal spectrogram* is constrained to fit the observation, $\mathbf{X}^{\text{cor}}$, at the reliable positions, while the *signal itself* follows the estimated AR model. To formalize this modification, note that the error in (3) can be written as $\mathbf{e} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{(N+p) \times N}$ is a Toeplitz matrix

composed using the (fixed) coefficients $\mathbf{a}$. We can then pose the proposed modification of step 2 as

$$\mathbf{x}^{(i+1)} = \arg\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{A}\mathbf{x}\|^2 + \iota_{\mathit{\Gamma}_{\mathrm{TF}}}(\mathbf{L}\mathbf{x}), \tag{4}$$

where $\mathbf{L}$ is the STFT and $\iota_{\mathit{\Gamma}_{\mathrm{TF}}}$ denotes the indicator function of the convex set $\mathit{\Gamma}_{\mathrm{TF}}$ of feasible spectrograms.[3] As such, the problem (4) is convex and it is solvable using standard numerical methods. We propose to employ the alternating direction method of multipliers (ADMM), which includes two principal steps [28, Sec. 3.1.1]:

$$\bar{\mathbf{x}}^{(k+1)} = \arg\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{A}\mathbf{x}\|^2 + \tfrac{\rho}{2}\|\mathbf{L}\mathbf{x} - \mathbf{z}^{(k)} + \mathbf{u}^{(k)}\|^2, \tag{5a}$$

$$\mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}} \iota_{\mathit{\Gamma}_{\mathrm{TF}}}(\mathbf{z}) + \tfrac{\rho}{2}\|\mathbf{L}\bar{\mathbf{x}}^{(k+1)} - \mathbf{z} + \mathbf{u}^{(k)}\|^2. \tag{5b}$$

The whole algorithm is summarized in Alg. 1. For the simplification of (5a), we have posed the assumption[4] $\mathbf{L}^*\mathbf{L} = \mathbf{I}$, leading to $\bar{\mathbf{x}}^{(k+1)} = \operatorname{prox}_{\frac{1}{2}\|\cdot\|^2}(\mathbf{L}^*(\mathbf{z}^{(k)} - \mathbf{u}^{(k)}))$, see [29, Rem. 3], which then corresponds to line 7 of Alg. 1 [30, Table 1]. Similarly, (5b) boils down to line 8 of the algorithm; this is computed via (2). The hyperparameters of the algorithm are the number of outer and inner iterations (I and K, respectively), and the step size $\rho > 0$.

Algorithm 1: Janssen-TF-ADMM

```
1  initialize x^(1)
2  for i = 1, ..., I do
       // AR model update
3      a^(i) = lpc(x^(i))
       // signal update using ADMM
4      form the Toeplitz matrix A from a^(i)
5      z^(1) = L x^(i), u^(1) = 0
6      for k = 1, ..., K do
7          x̄^(k+1) = (I + (1/ρ) A^T A)^(-1) L^* (z^(k) - u^(k))
8          z^(k+1) = proj_{Γ_TF}(L x̄^(k+1) + u^(k))
9          u^(k+1) = u^(k) + L x̄^(k+1) - z^(k+1)
10     end
11     x^(i+1) = x̄^(K+1)
12 end
```

Note that several other algorithms can be used to solve (4) numerically, such as the primal–dual algorithm [31] or the Douglas–Rachford or the forward–backward algorithms [30], combined with the approximal operator of the function $\iota_{\mathit{\Gamma}_{\mathrm{TF}}}(\mathbf{L}\,\cdot)$ [32]. However, the variant with ADMM described above, denoted Janssen-TF-ADMM, exhibited the best performance out of six algorithms in our preliminary testing.

V. Experiments

A. Data, setup and metrics

For our first experiment, we adopt the audio dataset[5] used for the evaluation of DPAI [18]. It consists of 6 music and 2 speech recordings of 5 seconds in length, sampled at 16 kHz. Only minor modifications were made to the original distribution of gaps (i.e., the placement of the missing spectrogram columns), such that each test excerpt includes 5 gaps of the same length. The gap lengths range from 1 to 6 missing columns. In turn, the entire test set comprises $8 \times 6 = 48$ corrupt spectrograms. Even though the dataset is relatively small and includes only a few types of audio signals, it provides a fair comparison of the proposed method with DPAI—the choice allows the use of exactly the same U-net architecture of DPAI from [18]. In particular, we chose their "best2" setting, and 5000 epochs of the Adam optimizer.

To validate the results on a larger dataset, we hand-picked 60 music recordings of various complexity from the IRMAS dataset [33], [34]. To enable the use of DPAI, the recordings were subsampled to 16 kHz and shortened to 5 seconds.

The STFT (i.e., the operator $\mathbf{L}$) utilizes a 2048-sample-long Hann window (corresponding to 128 ms[6]), with 75% overlap and 2048 frequency channels; this setting is the default for the time-frequency transform of the librosa package[7], as used in [18], and ensures that $\mathbf{L}$ forms a Parseval tight frame.

We compare the proposed method Janssen-TF-ADMM with three baseline methods:

- DPAI reconstruction with context,
- DPAI reconstruction without context,
- gap-wise Janssen method in the time domain [12].

The gap-wise method is a recent twist to the original Janssen algorithm [2], which is shown in [12] to be the state-of-the-art time-domain inpainter. As explained in Sec. II, the time-domain Janssen method is not directly comparable with the first two. We use it such that we transform the corrupted spectrogram $\mathbf{X}^{\mathrm{cor}}$ to the time-domain and zero-out all samples affected by the zero columns of $\mathbf{X}^{\mathrm{cor}}$.

To objectively evaluate the reconstruction quality, we assess the difference between the inpainted signal $\hat{\mathbf{x}}$ and the clean reference $\mathbf{x}^{\mathrm{orig}}$ using two standard metrics:

1) signal-to-noise ratio (SNR), defined in decibels as $\mathrm{SNR}(\hat{\mathbf{x}}, \mathbf{x}^{\mathrm{orig}}) = 10\log_{10}\frac{\|\mathbf{x}^{\mathrm{orig}}\|^2}{\|\mathbf{x}^{\mathrm{orig}} - \hat{\mathbf{x}}\|^2}$ [3],
2) objective difference grade (ODG) computed using the PEMO-Q [35], which takes psychoacoustics into account and predicts the subjective evaluation of the difference between $\hat{\mathbf{x}}$ and $\mathbf{x}^{\mathrm{orig}}$ on the scale from -4 to 0 (from "very annoying" to "imperceptible").

In addition, we run

3) a subjective listening test containing a limited number of excerpts. The MUSHRA-type test [36] was performed, which includes the reference signal, the competing reconstructions, and finally the hidden anchor, i.e., the signal reconstructed from the spectrogram with gaps. The participants were asked to score the perceived similarity of the signals with the reference signal on a continuous

[3] The original Janssen algorithm minimizes the function $\frac{1}{2}\|\mathbf{A}\mathbf{x}\| + \iota_{\mathit{\Gamma}_{\mathrm{T}}}(\mathbf{x})$ in subproblem 2.

[4] This assumption is met for instance if $\mathbf{L}$ corresponds to a Parseval tight frame, which can be ensured by a suitable practical setting of the STFT.

[5] https://github.com/fmiotello/dpai/

[6] This exceeds the standard window length for speech processing; however, a change of this hyperparameter would cause the need for changing the DPAI architecture.

[7] https://librosa.org/doc/0.10.2/generated/librosa.stft.html

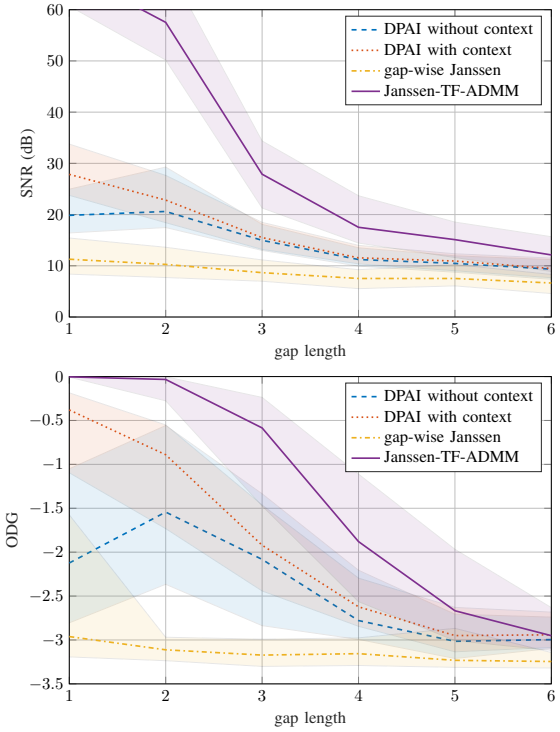

Fig. 1. Comparison of the inpainting methods using objective metrics – SNR (top) and ODG (bottom). The plots show results averaged over the 8 test signals together with the bootstrap interval estimates of the mean values at the $\alpha = 5\%$ significance level [37, p. 160]. If the intervals do not overlap, it may be concluded that the difference of the means is statistically significant.

scale from 0 to 100, with the help of the webMUSHRA environment[8]. The test was run in a quiet music studio, using a professional sound card and headphones. The listening conditions were identical for all the participants. Following the recommendation [36], 11 assessors out of 18 were selected through post-screening.

B. Results

The objective comparison on the DPAI dataset is shown in Fig. 1. The first observation is that the time-domain-based Janssen method [12] is not competitive with the other approaches. While the average performance of the DPAI variants mostly lies within the confidence intervals of each other, suggesting limited significance of the difference, the proposed Janssen-TF-ADMM method largely outperforms the competitors in both SNR and ODG (except the longest gaps).

To support the objective evaluation on the first dataset, Fig. 2 presents the results of the subjective listening test. Its results correlate well with the objective evaluation; in particular, Janssen-TF-ADMM is superior to the other methods. Moreover, this result is statistically significant in terms of medians. The DPAI variants scored almost the same.

Finally, Fig. 3 shows the evaluation on the IRMAS subset comprising 60 signals. This evaluation supports the previous analysis and leads to the conclusion that the proposed method

[8]https://github.com/audiolabs/webMUSHRA/

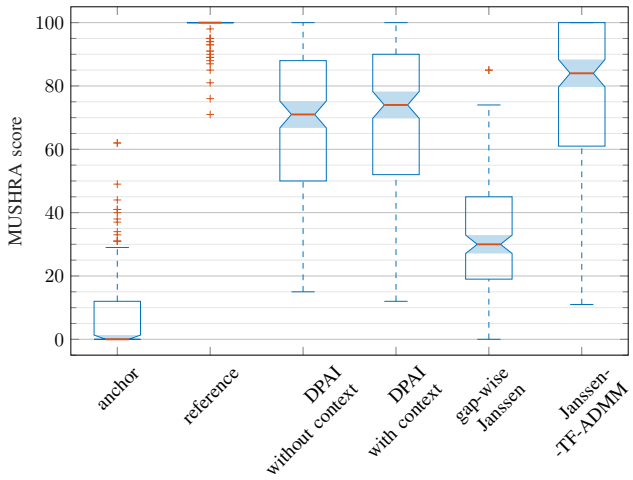

Fig. 2. A boxplot showing the distribution of scores in the listening test. The individual boxes span from the 25th to the 75th percentile of the recorded scores. The notches (filled areas) around the medians (orange lines) are constructed such that boxes whose notches do not overlap have different medians at the 5% statistical significance level.

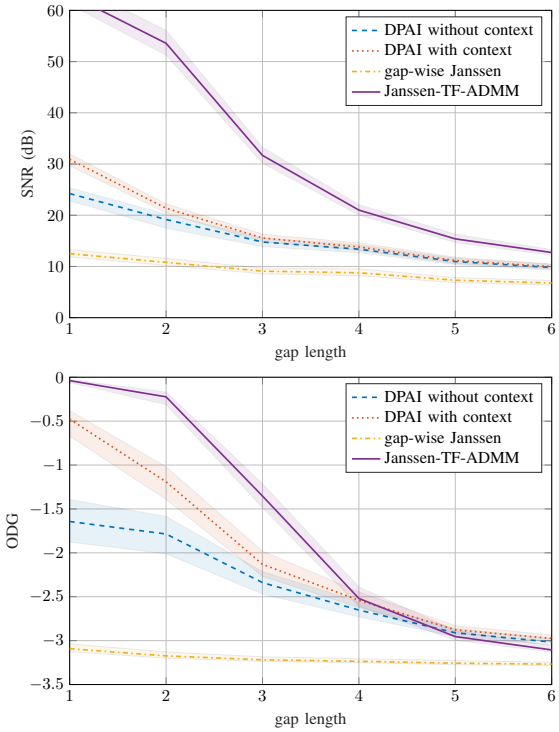

Fig. 3. SNR and ODG results on the IRMAS dataset.

is superior to the others. For shorter gaps, DPAI with context is preferable over DPAI not exploiting context.

Regarding computational demands, a single run of DPAI takes ca 19 minutes on the NVIDIA Tesla V100S GPU with 32 GB of memory, regardless of the number of missing spectrogram columns. Reconstructing a single signal with Janssen-TF-ADMM takes 10 to 20 minutes, proportional to the number of columns; this performance corresponds to a PC with Intel Core i7 3.40 GHz processor, 32 GB RAM.

VI. Conclusion and Outlook

The paper has shown that the proposed method of spectrogram inpainting, Janssen-TF, performs significantly better than the recently introduced DPAI algorithm, which is based on the deep prior idea. The conclusion has been certified both by objective and by subjective tests. In the future, it might be interesting to use the same adaptation strategy with other methods for the time-domain inpainting, such as the PHAIN regularized algorithm [13].

Matlab source codes for Janssen-TF and other supplementary material are publicly available at https://github.com/rajmic/spectrogram-inpainting.

References

[1] 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, 1996.

[2] A. J. E. M. Janssen, R. N. J. Veldhuis, and L. B. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, no. 2, pp. 317–330, Apr. 1986.

[3] Amir Adler, Valentin Emiya, Maria G. Jafari, Michael Elad, Rémi Gribonval, and Mark D. Plumbley, "Audio Inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

[4] Ondřej Mokrý and Pavel Rajmic, "Audio inpainting: Revisited and reweighted," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2906–2918, 2020.

[5] Ondřej Mokrý, Pavel Záviška, Pavel Rajmic, and Vítězslav Veselý, "Introducing SPAIN (SParse Audio INpainter)," in *2019 27th European Signal Processing Conference (EUSIPCO)*. 2019, IEEE.

[6] Ondřej Mokrý, Paul Magron, Thomas Oberlin, and Cédric Févotte, "Algorithms for audio inpainting based on probabilistic nonnegative matrix factorization," *Signal Processing*, p. 10, Dec. 2022.

[7] Yuval Bahat, Yoav Y. Schechner, and Michael Elad, "Self-content-based audio inpainting," *Signal Processing*, vol. 111, no. 0, pp. 61–72, 2015.

[8] Ichrak Toumi and Valentin Emiya, "Sparse non-local similarity modeling for audio inpainting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2018, IEEE.

[9] Nathanael Perraudin, Nicki Holighaus, Piotr Majdak, and Peter Balazs, "Inpainting of long audio segments with similarity graphs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[10] Gal Greshler, Tamar Shaham, and Tomer Michaeli, "Catch-A-Waveform: Learning to generate audio from a single short example," in *Advances in Neural Information Processing Systems*, M. Ranzato et al., Eds. 2021, vol. 34, pp. 20916–20928, Curran Associates, Inc.

[11] Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki, "Solving audio inverse problems with a diffusion model," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, IEEE.

[12] Ondřej Mokrý and Pavel Rajmic, "Tweaking autoregressive methods for inpainting of gaps in audio signals," Accepted to EUSIPCO 2025. https://arxiv.org/abs/2403.04433

[13] Tomoro Tanaka, Kohei Yatabe, and Yasuhiro Oikawa, "PHAIN: Audio inpainting via phase-aware optimization with instantaneous frequency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4471–4485, 2024.

[14] Marie Daňková, Pavel Rajmic, and Radovan Jiřík, "Acceleration of perfusion MRI using locally low-rank plus sparse model," in *Latent Variable Analysis and Signal Separation*, Liberec, 2015, pp. 514–521, Springer.

[15] P. Rajmic, Z. Koldovský, and M. Daňková, "Fast reconstruction of sparse relative impulse responses via second-order cone programming," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'17)*, New Paltz, NY, USA, Oct. 2017.

[16] Andrés Marafioti, Nathanaël Perraudin, Nicki Holighaus, and Piotr Majdak, "A context encoder for audio inpainting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2362–2372, Dec. 2019.

[17] Andres Marafioti, Piotr Majdak, Nicki Holighaus, and Nathanaël Perraudin, "GACELA: A generative adversarial context encoder for long audio inpainting of music," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 120–131, Jan. 2021.

[18] Federico Miotello, Mirco Pezzoli, Luca Comanducci, Fabio Antonacci, and Augusto Sarti, "Deep Prior-Based Audio Inpainting Using Multi-Resolution Harmonic Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[19] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Deep image prior," *International Journal of Computer Vision*, vol. 128, no. 12, 2020.

[20] P. Balazs, M. Dörfler, M. Kowalski, and B. Torrésani, "Adapted and adaptive linear time-frequency representations: a synthesis point of view," *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 20–31, Nov. 2013.

[21] Pavel Rajmic, Hana Bartlová, Zdeněk Průša, and Nicki Holighaus, "Acceleration of audio inpainting by support restriction," in *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. Oct. 2015, pp. 325–329, IEEE.

[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[23] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Spain, 2020, pp. 6199–6203, IEEE.

[24] Eloi Moliner, Michal Švento, Alec Wright, Lauri Juvela, Pavel Rajmic, and Vesa Välimäki, "Unsupervised estimation of nonlinear audio effects: Comparing diffusion-based and adversarial approaches," 2025.

[25] Nabil Ibtehaz and M. Sohel Rahman, "Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[26] Hirotoshi Takeuchi, Kunio Kashino, Yasunori Ohishi, and Hiroshi Saruwatari, "Harmonic lowering for accelerating harmonic convolution for audio signals," in *Proc. Interspeech*, 10 2020, pp. 185–189.

[27] Peter J. Brockwell and Richard A. Davis, *Time Series: Theory and Methods*, Springer series in statistics. Springer, 2 edition, 2006.

[28] Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers.," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 122, 2011.

[29] Pavel Záviška, Ondřej Mokrý, and Pavel Rajmic, "S-SPADE Done Right: Detailed Study of the Sparse Audio Declipper Algorithms," techreport, Brno University of Technology, 2018. https://arxiv.org/abs/1809.09847

[30] Patrick L. Combettes and Jean-Christophe Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, vol. 49, pp. 185–212, 2011.

[31] Antonin Chambolle and Thomas Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[32] Ondřej Mokrý and Pavel Rajmic, "Approximal operator with application to audio inpainting," *Signal Processing*, vol. 179, pp. 107807, 2021.

[33] Juan J. Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals," *International Society for Music Information Retrieval Conference*, pp. 559–564, 2012.

[34] Juan J. Bosch, Ferdinand Fuhrmann, and Perfecto Herrera, "IRMAS: A dataset for instrument recognition in musical audio signals," 2014. DOI 10.5281/ZENODO.1290750

[35] R. Huber and B. Kollmeier, "PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio Speech Language Proc.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[36] "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[37] Bradley Efron and Robert J. Tibshirani, *An Introduction to the Bootstrap*, Chapmann & Hall, New York, 1994.