On the Use of Autoregressive Methods for Audio Inpainting

Ondřej Mokrý
Brno University of Technology,
FEEC, Dept. of Telecommunications,
Technická 12, 616 00 Brno, Czech Republic
E-mail: ondrej.mokry@vut.cz

Pavel Rajmic
Brno University of Technology,
FEEC, Dept. of Telecommunications,
Technická 12, 616 00 Brno, Czech Republic
E-mail: pavel.rajmic@vut.cz

arXiv:2403.04433v1 [eess.AS] 7 Mar 2024

***Abstract*—The paper presents an evaluation of popular audio inpainting methods based on autoregressive modelling, namely, extrapolation-based and Janssen methods. A novel variant of the Janssen method suitable for gap inpainting is also proposed. The main differences between the particular popular approaches are pointed out, and a mid-scale computational experiment is presented. The results demonstrate the importance of the choice of the AR model estimator and the suitability of the new gap-wise Janssen method.**

***Index Terms*—audio, autoregression, inpainting, interpolation**

I. Introduction

Audio inpainting is a challenging signal processing task, where missing parts of an audio signal have to be completed. For a human listener, the result should be as pleasant as possible and ideally not noticeable. Previously proposed audio inpainting solutions cover a wide range of approaches, from autoregressive modelling [1]–[5], through optimisation methods [6]–[10] to deep learning [11]–[13] or graph-based methods [14].

For signal gaps up to ca 80 milliseconds, the iterative method of Janssen et al. [1] proposed in 1986 constantly ranks among the best, according to numerous studies [7], [10], [15], [16]. The extrapolation methods [3]–[5], [17] are non-iterative and utilise a twofold extrapolation (from left to right and right to left) while the two particular solutions are blended together using a crossfading scheme. Such an approach belongs among the most popular, perhaps thanks to its simplicity and speed, and is actually used in the Matlab function fillgaps. The patented method of Etter [2] considers the just mentioned approach suboptimal and proposes to join the two directions of extrapolation using a single optimisation criterion.

Besides slight variances in how to *model the signal*, different algorithms for *estimation of the coefficients* are available [19], which will be discussed in detail further on.

Unfortunately, a thorough, systematic comparison of autoregressive approaches seems to be absent in the literature. In this paper, we review the principle of autoregression-based methods, point out main differences between the particular popular approaches, and present a mid-scale computational experiment on an audio inpainting dataset. Last but not least, we also propose a new variant not present in the literature and examine its performance compared to known approaches.

The work was supported by the Czech Science Foundation (GAČR) Project No. 23-07294S.

II. Modelling audio as an autoregressive process

An autoregressive (AR) process of order p is a discrete-time stochastic process $\{X_n \mid n = 0, \pm 1, \pm 2, \dots\}$ defined by the relation

$$X_n + a_2 X_{n-1} + \cdots + a_{p+1} X_{n-p} = E_n \quad \text{for all } n, \tag{1}$$

where $\{E_n\}$ is white noise process with a zero-mean and variance $\sigma^2 > 0$ [19, Def. 3.1.2]. Expression (1) can be interpreted as passing the excitation noise $\{E_n\}$ through an all-pole filter with coefficients $a_1, a_2, \dots, a_{p+1}$, where $a_1 = 1$ [20, Ch. 4]. The order p defines the range of indexes that determine the output in the current time instance. As such, the frequency resolution increases with increasing p.

For a particular observation of the process, $\mathbf{x} \in \mathbb{R}^N$, (1) can be written as

$$\sum_{i=1}^{p+1} a_i x_{n+1-i} = e_n, \quad a_1 = 1, \ n = 1, \dots, N+p. \tag{2}$$

Vector $\mathbf{x}$ is zero-padded to its new length $N + p$, such that (2) is correctly defined. The above is closely related to the notion of convolution, and it is actually in accordance with the convention of the lpc function in Matlab.

From the perspective of model fitting, the vector $\mathbf{e} \in \mathbb{R}^{N+p}$ is called *residual error* [21, Sec. 8.2.2]. Given the observed signal $\mathbf{x}$ and the order p, the coefficients of the AR model are usually estimated via the optimization problem

$$\arg\min_{\mathbf{a}} \frac{1}{2} \|\mathbf{e}(\mathbf{a}, \mathbf{x})\|^2 \tag{3}$$

where we denote $\mathbf{a} = [1, a_2, a_3, \dots, a_{p+1}]^\top$, and $\mathbf{e} = [e_1, e_2, \dots, e_{N+p}]^\top$ is defined in (2) as a function of both the signal $\mathbf{x}$ and the coefficients $\mathbf{a}$. This problem can be effectively solved using, for example, the Levinson–Durbin algorithm [22], [23]. Using the estimate of autocorrelation naturally imposes the notion of the autocorrelation algorithm; however, it is commonly referred to simply as the LPC algorithm, where LPC stands for linear prediction coefficients.

In contrast to LPC, Burg algorithm [24] for the estimation of the AR parameters involves an extra assumption that the *same* parameters should model both the signal $\mathbf{x}$ and its version flipped in time. In effect, this results in another quadratic term that extends (3). Kauppinen and Roth prefer the Burg algorithm for audio signal extrapolation [4, Sec. 4.2] since the underlying all-pole filter obtained this way is stable [18, Sec. 12.3.3]. A frequency-warped Burg algorithm has been proposed [17] that allows focus on specified spectral bands, however, the effect of warping is comparable to increasing the model order p in the non-warped case [17, Sec. 5].

Additional terms can be optionally appended to the objective (3) that *regularise* either the reconstructed signal or the AR coefficients, as proposed in [25], [26]. However, such an approach is not explored satisfactorily, and we do not include it in the present treatment. The role and effect of regularisation in autoregressive modelling will be treated in a separate contribution of Mokrý and Rajmic, currently in preparation.

III. Audio inpainting using autoregression

To formalize the problem of audio inpainting, assume that the observed signal $\mathbf{x} = [x_1, \dots, x_N]^\top \in \mathbb{R}^N$ consists of reliable samples identified by the set of indices $M \subset \{1, \dots, N\}$, and vacant samples at positions $\overline{M} = \{1, \dots, N\} \setminus M$. The goal of inpainting is to estimate the missing samples at positions $\overline{M}$. In the so-called consistent case, the samples at positions M are meant to be preserved, i.e., any candidate solution $\hat{\mathbf{x}}$ to the inpainting problem should satisfy $\hat{x}_n = x_n$ for all $n \in M$.

In this work, the focus is on the practical scenario where the signal contains several gaps, i.e., segments of consecutive lost samples, surrounded by an intact context. For this scenario, two approaches based on AR modelling are applicable.

First, the extrapolation-based method fits two independent sets of AR parameters for each gap, one for the left context and one for the right context of the gap. These coefficients are then used to extrapolate (predict) both contexts inside the gap, and the forward- and backward-extrapolated signals are then cross-faded using a predefined suitable function. Although numerous fading options are possible (see, for instance, [3, Sec. 4.2]), we resort to the raised cosine function [2, Sec. II.].

Second, the Janssen method is based upon alternating estimation of the AR model for a signal frame and the missing samples in this frame [1]. The method iteratively updates the AR coefficients of a signal frame (the current estimate of the missing samples being fixed) and the missing samples (with fixed estimate of the model parameters). As such, it solves[1] a problem similar to (3) where both $\mathbf{a}$ and $\mathbf{x}$ are variables, and $\mathbf{x}$ is constrained to stick to the reliable part of a signal frame. Since the AR model is most suitable for stationary signals, the method is usually applied independently in temporal signal frames. These frames typically overlap in time, and they can be extracted using rectangular window (as mentioned in [6, Tab. III]), but numerous other options are readily available.

Finally, we propose a novel use of the Janssen method, which, instead of overlapping frames, treats each gap in the signal separately (hence the name gap-wise Janssen method). This means that a single AR model is fitted to *both* left and right contexts of the gap.[2] Such a way of using the context is similar with the extrapolation-based method, but the AR coefficients (shared by both contexts) are estimated as in Janssen method. Note that this approach first appeared in [10] with hyperparameters (the context length, model order and estimation algorithm) chosen intuitively. Apparently, the proposed approach cannot be used in scenarios with random positions of missing samples, in contrast to the above window-based version.

[1] Such a problem is non-convex. Even though each Janssen iteration is guaranteed to decrease the objective, it might fail to find the global optimum.

IV. Experiments & Results

The test signals are 9 recordings of individual musical instruments, taken from the EBU SQAM database [27]. They are sampled at 44.1 kHz and cropped to a length of around 7 seconds. We chose solo instruments since AR models are expected to perform well on them; a sum of multiple AR processes generally may not be an AR process of order reasonably comparable with the individual signal orders. Therefore, the more complex the signal, the less suitable it is for AR modelling.

To simulate degradation, we consider gap lengths from 10 ms up to 80 ms, and create 10 gaps in each signal at pseudorandom locations.[3] From AR-based methods, we use all three aforementioned approaches: The extrapolation-based method and gap-wise Janssen are applied with fixed context length 4096 samples (approx. 93 ms). The frame-wise Janssen uses window length 4096 samples and 3 window shapes: rectangular, Hann, and Tukey (see, e.g., [28, Sec. V]). All methods are applied with variable model order p and either Burg or LPC algorithm to fit the AR model.

The quality of reconstructed audio is assessed using the signal-to-distortion ratio (SDR). For the reference (i.e., undegraded) signal $\mathbf{y}$ and the reconstruction $\hat{\mathbf{x}}$, SDR in decibels is computed as $\mathrm{SDR}(\mathbf{y}, \hat{\mathbf{x}}) = 10 \log_{10} \frac{\|\mathbf{y}\|^2}{\|\mathbf{y} - \hat{\mathbf{x}}\|^2}$. In our case, the SDR is only computed in the inpainted sections of the signal.

The perceived quality of the signal is evaluated using the objective metric PEMO-Q [29], which predicts the subjective difference of the signals $\mathbf{y}$ and $\hat{\mathbf{x}}$ in terms of the objective difference grade (ODG), ranging from -4 (very annoying) to 0 (imperceptible difference). An alternative choice, which is common in other audio processing fields, is PEAQ [30], [31]. However, this metric is not decisive enough in the case of gap inpainting [32].

A. Effect of the estimator

First, we evaluate the performance of the inpainting methods depending on the estimator of the AR model. Hence, for each test instance, two versions of each inpainting method are run,

[2] If, furthermore, the Burg algorithm is used to estimate the AR coefficients, it means that a single AR model is assumed not only for the whole gap context, but also for its flipped version.

[3] The used signals are available in the repository https://github.com/ondrejmokry/TestSignals.

varying in the use of either LPC or Burg algorithm to estimate the AR parameters.

The results are presented in Fig. 1. The overall distribution of the results in the scatter plots indicates that in terms of SDR, the Burg algorithm is clearly favourable in the case of the extrapolation-based inpainting. For the considered iterative algorithms, the results indicate minor preference of the LPC. In terms of the ODG, the observations are much more pronounced. Besides extrapolation-based inpainting, the preferability of Burg algorithm is indicated also in the gap-wise Janssen algorithm. Notably, a conclusion in case of the frame-wise Janssen algorithm depends on the model order and the selected window. With the Hann or Tukey window, inpainting results appear to depend on the chosen estimator only for large model orders (and the LPC scores better in this case). With the rectangular window, the dependence on the model order is clearly amplified. Furthermore, a large portion of results in this case implies that the Burg algorithm is a better choice if the model order is low, and vice versa.

To assess statistical significance of the results, Wilcoxon signed rank test[4] is performed on the paired results obtained using LPC and Burg algorithm. The p-values, for each method and model order, are shown in Tab. I. If the p-value is lower than the chosen significance level α, it indicates the rejection of the null hypothesis that the Burg algorithm and LPC lead to results having an identical median, with the alternate hypothesis that the Burg algorithm leads to a higher median.

B. Effect of the model order

The AR model order p plays a crucial role in the modelling and also significantly affects the inpainting results, as demonstrated in Fig. 2. Note that Fig. 2 plots the same data as Fig. 1, but with a focus on the effect of the model order and the gap length.

Similarly to the study of preferences between LPC and Burg algorithm, the model order affects the results differently for different inpainting methods. A clear scheme is observed in the case of the extrapolation-based and gap-wise Janssen method, where increasing the model order up to $p = 2048$ results in both higher SDR and higher ODG. However, the order $p = 3072$ does not further increase the resulting quality, which we attribute to insufficient context length for fitting an AR model of such a high order. For frame-wise Janssen, the results are further affected by the chosen window shape: the best results are scored using model order $p = 512$ with Tukey or rectangular window and $p = 1024$ with Hann window.

Furthermore, Fig. 2 reveals that the observed phenomena do not largely depend on the length of the gap.

C. Comparison with other methods

To provide a context for the results of AR-based methods, selected variants are compared with the methods that belong to the state-of-the-art in optimisation-based audio inpainting, namely A-SPAIN [7] and A-SPAIN-MOD [15]. The results in terms of SDR and ODG are presented in Fig. 3. The most significant observation is the dominance of the extrapolation-based and gap-wise Janssen methods, especially in gaps longer than 50 ms.

[4] https://www.mathworks.com/help/stats/signrank.html

V. Conclusion

The paper presented an evaluation of popular audio inpainting methods based on autoregressive modelling, both using extrapolation and iterative estimation. In addition, a novel gap-wise approach to the Janssen method was proposed. The experiments demonstrated the importance of the choice of the AR model estimator (i.e. choosing the LPC or Burg algorithm) and the model order. The concluding test revealed that the gap-wise Janssen method (using Burg algorithm) is recommended as an autoregressive reference for future tests on inpainting middle-length gaps.

If computational speed is an important criterion, note that for all approaches, the computational load is proportional both to the order of the AR model and to the gap length. Moreover, the Burg algorithm is more demanding compared to LPC. From the perspective of computational time, the extrapolation-based approach is preferred (elapsed times are up to around 0.15 s per signal with $p = 2048$, while the gap-wise Janssen reaches up to 11.5 s per signal with $p = 1024$ and up to 16 s with $p = 2048$; both using the Burg algorithm).

The Matlab codes for the methods discussed in the present paper are available at https://github.com/ondrejmokry/InpaintingAutoregressive.

References

[1] 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.

[2] W. Etter, "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Trans. Sig. Proc.*, vol. 44, no. 5, pp. 1124–1135, 1996.

[3] I. Kauppinen and J. Kauppinen, "Reconstruction method for missing or damaged long portions in audio signal," *Journal of the Audio Engineering Society*, vol. 50, no. 7/8, pp. 594–602, 2002.

[4] I. Kauppinen and K. Roth, "Audio signal extrapolation – Theory and applications," in *Proceedings of the 5th International Conference on Digital Audio Effects*, Hamburg, Germany, Sep. 2002, pp. 105–110.

[5] P. A. A. Esquef, V. Välimäki, K. Roth, and I. Kauppinen, "Interpolation of long gaps in audio signals using the warped Burg's method," in *Proceedings of the 6th International Conference on Digital Audio Effects*, London, UK, Sep. 2003, pp. 18–23.

[6] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

TABLE I
THE P-VALUES OF THE WILCOXON SIGNED RANK TEST. VALUES EXCEEDING THE COMMON LEVEL $\alpha = 0.05$ ARE IN BOLD.

	comparing SDR (dB)					comparing ODG				
method / model order	256	512	1024	2048	3072	256	512	1024	2048	3072
extrapolation-based	$1.1 \cdot 10^{-21}$	$2.1 \cdot 10^{-22}$	$2.1 \cdot 10^{-22}$	$2.1 \cdot 10^{-22}$	$3.6 \cdot 10^{-20}$	$2.1 \cdot 10^{-22}$	$2.1 \cdot 10^{-22}$	$2.1 \cdot 10^{-22}$	$2.1 \cdot 10^{-22}$	$5.3 \cdot 10^{-21}$
Janssen, gap-wise	$2.2 \cdot 10^{-09}$	$3.9 \cdot 10^{-13}$	$5.5 \cdot 10^{-10}$	$3.1 \cdot 10^{-02}$	$\mathbf{9.7 \cdot 10^{-01}}$	$4.2 \cdot 10^{-22}$	$2.1 \cdot 10^{-22}$	$1.8 \cdot 10^{-18}$	$1.4 \cdot 10^{-08}$	$1.7 \cdot 10^{-02}$
Janssen, Hann window	$7.8 \cdot 10^{-05}$	$6.2 \cdot 10^{-04}$	$\mathbf{9.0 \cdot 10^{-01}}$	$\mathbf{1.0 \cdot 10^{+00}}$	$\mathbf{1.0 \cdot 10^{+00}}$	$4.3 \cdot 10^{-03}$	$7.5 \cdot 10^{-03}$	$\mathbf{9.5 \cdot 10^{-01}}$	$\mathbf{1.0 \cdot 10^{+00}}$	$\mathbf{1.0 \cdot 10^{+00}}$
Janssen, Tukey window	$2.6 \cdot 10^{-03}$	$\mathbf{1.3 \cdot 10^{-01}}$	$\mathbf{1.0 \cdot 10^{+00}}$	$\mathbf{1.0 \cdot 10^{+00}}$	$\mathbf{1.0 \cdot 10^{+00}}$	$5.2 \cdot 10^{-01}$	$8.3 \cdot 10^{-02}$	$\mathbf{1.0 \cdot 10^{+00}}$	$\mathbf{1.0 \cdot 10^{+00}}$	$\mathbf{1.0 \cdot 10^{+00}}$
Janssen, rect. Window	$6.7 \cdot 10^{-11}$	$4.7 \cdot 10^{-06}$	$\mathbf{3.9 \cdot 10^{-01}}$	$\mathbf{9.9 \cdot 10^{-01}}$	$\mathbf{1.0 \cdot 10^{+00}}$	$2.1 \cdot 10^{-22}$	$1.4 \cdot 10^{-15}$	$\mathbf{1.1 \cdot 10^{-01}}$	$\mathbf{4.2 \cdot 10^{-01}}$	$\mathbf{1.2 \cdot 10^{-01}}$

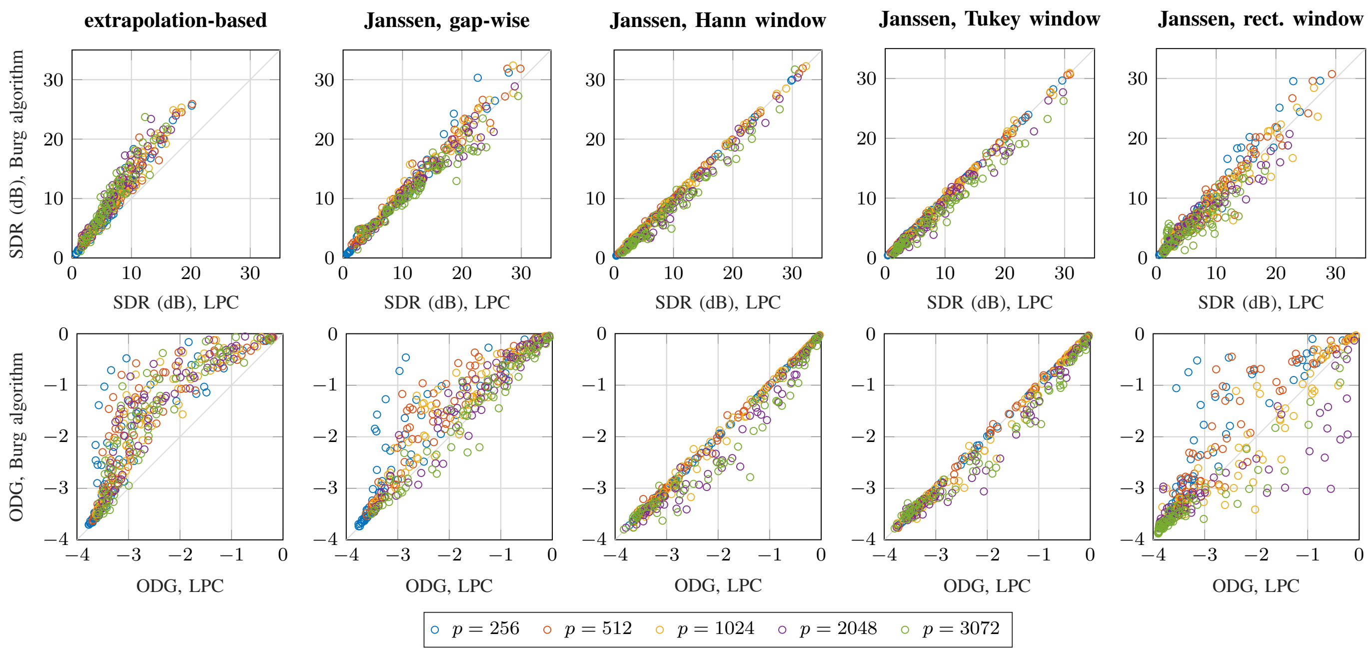

Fig. 1. Comparison of the estimators in terms of SDR (top row) and PEMO-Q ODG (bottom row). Per each inpainting method, the scatter plot shows the results using LPC vs. Burg algorithm to estimate the AR coefficients.

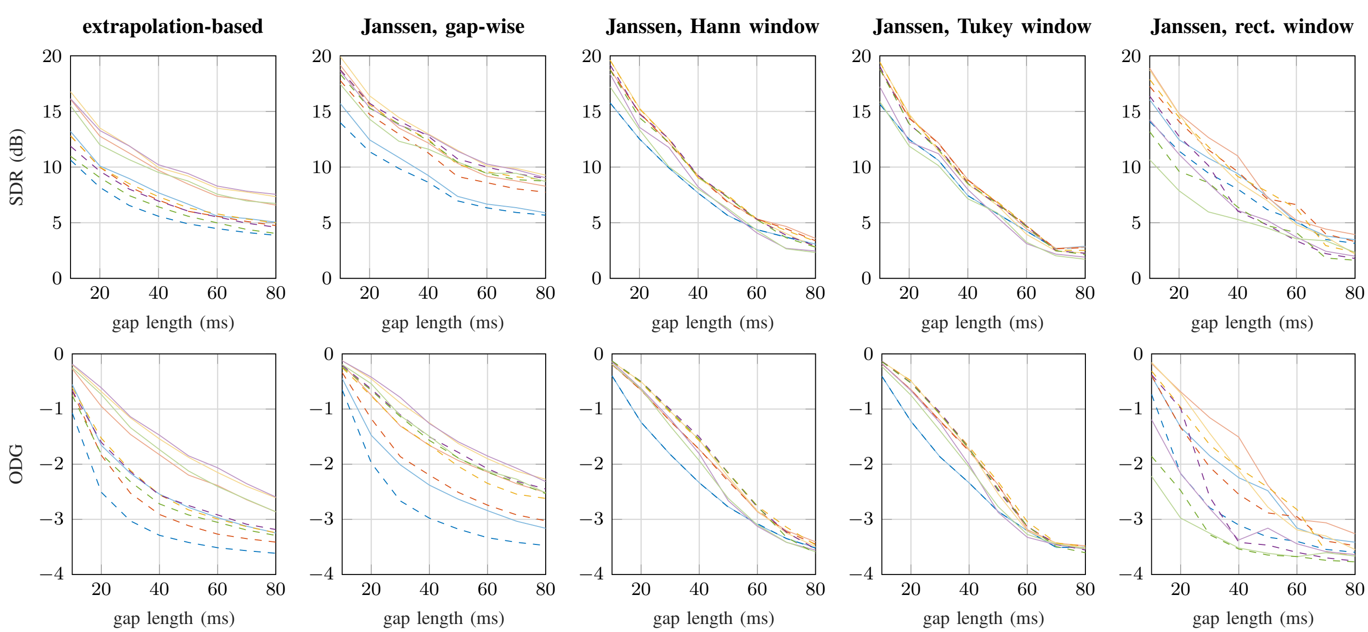

Fig. 2. Comparison of the model order choices in terms of SDR (top row) and PEMO-Q ODG (bottom row). Per each inpainting method, the plot shows the results using LPC (darker shade, dashed line) vs. Burg algorithm (lighter shade, solid line) to estimate the AR coefficients. Note that the same colour coding of the model order as in Fig. 1 applies.

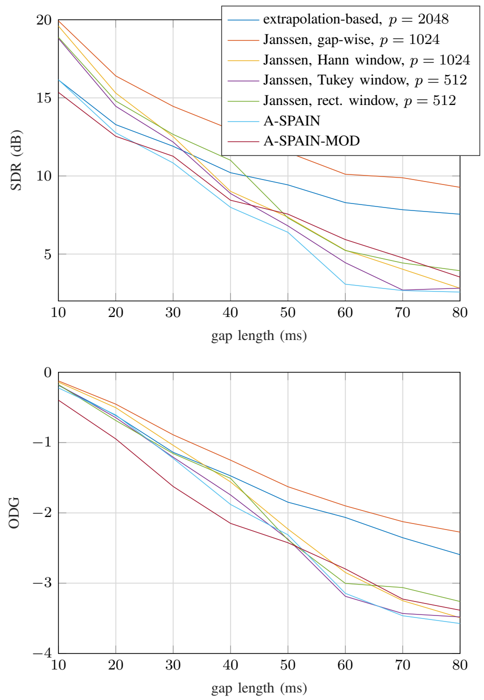

Fig. 3. Comparison of the AR-based methods with SPAIN in terms of SDR (top) and ODG (bottom), averaged over all signals. In this experiment, all AR-based methods used the Burg algorithm to estimate the coefficients.

[7] O. Mokrý, P. Záviška, P. Rajmic, and V. Veselý, "Introducing SPAIN (SParse Audio INpainter)," in *27th European Signal Processing Conference*, IEEE, 2019.

[8] O. Mokrý and P. Rajmic, "Audio inpainting: Revisited and reweighted," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 28, pp. 2906–2918, 2020.

[9] T. Tanaka, K. Yatabe, and Y. Oikawa, "Phase-aware audio inpainting based on instantaneous frequency," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2021.

[10] O. Mokrý, P. Magron, T. Oberlin, and C. Févotte, "Algorithms for audio inpainting based on probabilistic nonnegative matrix factorization," *Signal Proc.*, 2022.

[11] E. Moliner and V. Välimäki, "Diffusion-based audio inpainting," Tech. Rep., Sep. 2023. DOI: 10.48550/arXiv.2305.15266.

[12] F. Miotello, M. Pezzoli, L. Comanducci, F. Antonacci, and A. Sarti, "Deep prior-based audio inpainting using multi-resolution harmonic convolutional neural networks," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, 2023.

[13] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A context encoder for audio inpainting," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 27, no. 12, Dec. 2019.

[14] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs, "Inpainting of long audio segments with similarity graphs," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, 2018.

[15] G. Tauböck, S. Rajbamshi, and P. Balazs, "Dictionary learning for sparse audio inpainting," *IEEE J. of Selected Topics in Signal Processing*, vol. 15, no. 1, 2021.

[16] P. Záviška, P. Rajmic, and O. Mokrý, "Multiple Hankel matrix rank minimization for audio inpainting," in *2023 46th International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, Jul. 2023.

[17] K. Roth, I. Kauppinen, P. A. A. Esquef, and V. Välimäki, "Frequency warped Burg's method for AR-modeling," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, 2003.

[18] J. G. Proakis and D. G. Manolakis, *Digital signal processing*, T. Robbins, Ed. Prentice Hall, 1996.

[19] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. Springer, 2006.

[20] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. John Wiley & Sons, Inc., 2005.

[21] U. Zölzer, *DAFX: Digital Audio Effects*, 2nd edition, U. Zölzer, Ed. Wiley, 2011.

[22] J. Durbin, "The fitting of time-series models," *Revue de l'Institut International de Statistique*, vol. 28, no. 3, p. 233, 1960.

[23] N. Levinson, "The Wiener (root mean square) error criterion in filter design and prediction," *Journal of Mathematics and Physics*, vol. 25, no. 1-4, Apr. 1946.

[24] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford University, May 1975.

[25] B. D. Dufera, K. Eneman, and T. van Waterschoot, "Missing sample estimation based on high-order sparse linear prediction for audio signals," in *2018 26th European Signal Processing Conference*, IEEE, 2018.

[26] B. D. Dufera, E. Adugna, K. Eneman, and T. van Waterschoot, "Restoration of click degraded speech and music based on high order sparse linear prediction," in *2019 IEEE AFRICON*, IEEE, Sep. 2019.

[27] "EBU SQAM CD: Sound quality assessment material recordings for subjective tests." (2008), [Online]. Available: https://tech.ebu.ch/publications/sqamcd.

[28] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[29] 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, 2006.

[30] T. Thiede, W. C. Treurniet, R. Bitto, *et al.*, "PEAQ – The ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, Jan. 2000.

[31] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," McGill University, Tech. Rep., May 2002.

[32] O. Mokrý, "Modern optimization methods for interpolation of missing sections in audio signals," Ph.D. dissertation, Brno University of Technology, 2024.