Deep prior audio compression

Michal Švento
Dept. of Telecommunications, FEEC
Brno University of Technology
Brno, Czech Republic
michal.svento@vut.cz

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

***Abstract*—Audio compression is still an up-to-date topic because the demand for big data streams is rapidly increasing. Deep learning has brought up new algorithms that decrease bitrates with good perception quality. The novel approach in generative artificial intelligence is to produce new data from prior stored in network parameters, called a deep prior. The deep audio prior framework shows its success in various tasks such as inpainting, declipping, and bandwidth extension, but it has not been tested for compression. In this paper, we test this method with a pre-built network for inpainting. Our idea of compression is based on reducing the number of time-frequency coefficients in the spectrogram while allowing the reconstruction of the original signal with high quality.**

***Index Terms*—audio processing, deep learning, deep audio prior, compression**

I. Introduction

The problem of signal compression is as old as the communication itself [1]. Despite the increased speed of the transmission channels, user demands also follow this trend. Currently in audio, there is a demand for real-time applications, such as music shows that take place at multiple locations simultaneously that require a fast data stream and are especially sensitive to latency. On the contrary, some applications need better sound quality than low latency. This makes the topic of audio compression attractive and opens up various research directions for improvement.

The operation of a generic codec, which stands for coder–decoder, can be categorised into two key parts. The *encoder* compresses the original data into a smaller size while maintaining the quality of perception as much as possible. The fundamental component of the encoder is *quantiser* where compressed data are effectively encoded for transmission. The *decoder* reconstructs the received signal with minimal deviation from the original unprocessed signal.

The key properties of an audio codec are decoded quality (ability to restore compressed audio with minimal loss), compression efficiency (low bitrate), generation speed, and latency (minimum time to initialise the codec) [2]. Audio codecs are split into two variants: waveform codecs and parametric codecs. The parametric codecs saves storage in underlying features such as the time-frequency (TF) characteristic. Popular TF representations are the Short-time Fourier Transform (STFT), Constant-Q Transform (CQT), or Modified Discrete Cosine Transform (MDCT). These are used by most successful coders. The most popular audio codec is the MPEG layer 3 (mp3) as a part of the video codec [3]. Its efficiency is based mostly on psychoacoustical principles. The ear tract is a mechanical system and is imperfect. The codec suppresses information below the absolute threshold of hearing and exploits frequency masking. The subsequent technologies, Advanced Audio Codec (AAC) [4] and the open-source Vorbis [5], enhance the concepts of the mp3 and result in more efficient storage utilization. These codecs are significant for low bitrate, but the decoded quality is generally worse than waveform codecs.

In contrast, waveform codecs manipulate with raw audio signal. The simplest idea is to downsample waveform data and then interpolate the dropped samples. The first digital codecs were built upon this idea. Algorithms for coder are fast and this technique is still present [6]. The advent of deep learning has also affected the area of codecs. For example, SoundStream [7], Encodec [8] as end-to-end waveform codecs. They used residual vector quantisation to reduce bitrate, while utilising the loss of the HiFi-GAN vocoder [9] to ensure the fidelity of the decoded audio. The novel state-of-the-art combines the strengths of the parametric and waveform codec, APCodec [2].

Deep neural networks also open up quite different ways of approaching problems in signal processing and restoration. Instead of an end-to-end or parametric approach in audio, we can optimise the prior of data hidden in network layers. These methods have shown success in image restoration tasks, e.g. Deep Image Prior – DIP [10]. Additional applications were developed from DIP, including variations with audio, e.g. the Deep audio prior [11] and Deep prior audio inpainting (DPAI) [12]. They have shown success for a variety of different restoration problems. Compression is a task that has not yet been tested with deep prior networks to our best knowledge. The similarity of data (de)compression and inpainting problems motivates the use of DPAI as a decoder of subsampled, i.e. compressed, data [6].

This paper is organised as follows. Section II describes the tools used to solve inpainting problems with the deep prior neural network. Section III contains the setup of the

The work was supported by the Czech Science Foundation (GAČR) Project No. 23-07294S. The authors are grateful to NVIDIA for donation of the Titan XP graphic card, which has been used in this research. PhD study of Michal Švento is supervised by Pavel Rajmic.

experiment and the choice of TF masks. Sections IV and V summarise the results and future work.

II. Problem formulation

Assume a single channel audio signal $\mathbf{x} \in \mathbb{R}^L$ sampled at sampled at the sampling rate of f_s (Hz) where L is the number of samples. The TF characteristic of the signal is computed using the STFT. Conversely, given the TF characteristic, the time-domain signal can be synthesised using the inverse STFT (iSTFT). Both operations are linear and can be represented as the analysis operator $\mathcal{A}$ (STFT) and the synthesis operator $\mathcal{D}$ (iSTFT). The analysis operator $\mathcal{A} : \mathbb{R}^L \to \mathbb{C}^{M\times N}$ transforms the signal $\mathbf{x}$ into a complex matrix (spectrogram)

$$\mathbf{X} = \mathcal{A}(\mathbf{x}) \in \mathbb{C}^{M\times N}. \tag{1}$$

The number M expresses the number of frequency bins, and N is the number of time frames. The synthesis operator is a mapping $\mathcal{D} : \mathbb{C}^{M\times N} \to \mathbb{R}^L$. With an appropriate setting for the pair of $\mathcal{A}$ and $\mathcal{D}$ the composition of the analysis and synthesis is the identity [13], i.e. $\mathbf{x} = \mathcal{D}(\mathcal{A}(\mathbf{x}))$ for any signal in the time domain $\mathbf{x}$.

For the task of audio inpainting, the damage is usually considered in the time domain. However, there is an alternative approach that simulates loss in the TF domain [12], [14]. For a clean spectrogram $\mathbf{X} \in \mathbb{C}^{M\times N}$, the damaged spectrogram $\widetilde{\mathbf{X}}$ is obtained according to

$$\widetilde{\mathbf{X}} = \mathbf{X} \odot \mathbf{M}, \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{M\times N}$ is mask matrix and $\odot$ is Hadamard (element-wise) product. The matrix $\mathbf{M}$ is a binary mask with value one when the corresponding complex coefficient is preserved, and zero if its discarded. The authors of [12] used only masks for the inpainting task where the full columns are removed. Our implementation extends these masks beyond these presumptions, and we remove coefficients from the whole coefficient matrix. The exact process for the generation of masks will be described in Section III-B.

A. Deep Prior

We would like to use the Deep Prior neural network for the audio compression scenario. Most approaches train the network and get the knowledge about parameter distribution from learning. However, a significant amount of data knowledge about the sound is contained within the network architecture even without performing any training of the model parameters. This approach uses a completely untrained generative convolutional neural network (CNN).

We can write inverse audio problem as an energy minimisation task:

$$\widetilde{\mathbf{X}}^* = \arg\min_{\mathbf{X}} E(\mathbf{X} \odot \mathbf{M}; \widetilde{\mathbf{X}}) + R(\mathbf{X}), \tag{3}$$

where $E(\cdot)$ is task-dependent data fidelity term. For inpainting (compression) tasks, it is mostly the Mean Square Error (MSE) or ℓ_1 norm. The fidelity part ensures the solution will not deflect from the reliable part of the spectrogram. The second part $R(\cdot)$ is the regularisation term. It captures generic prior in the spectrogram, and it is usually hard to describe. In deep prior approach, the regularizer $R(\cdot)$ is replaced by the neural network $f_\theta(\cdot)$ with parameters θ which captures the implicit prior [10]:

$$\theta^* = \arg\min_{\theta} E(f_\theta(\mathbf{Z}); \widetilde{\mathbf{X}}), \quad \mathbf{X}^* = f_{\theta^*}(\mathbf{Z}), \tag{4}$$

where $\mathbf{Z} \sim \mathcal{N}(0, 1)$ is an initial random noise realization. The minimiser θ^* is obtained using well-known optimisers, in our case Adam [15].

Due to the fact that the result is a repaired spectrogram, for an audible result, we need to synthesise these complex coefficients. Result signal is

$$\mathbf{x}^* = \mathcal{D}(\mathbf{X}^*). \tag{5}$$

Despite using a neural network, no pre-trained models are used. Only damaged audiosignal transformed into a spectrogram, and a proper network architecture are needed for successful reconstruction. In deep prior approach we do not look for answers in the audio domain (end-to-end), but in the domain of network parameters. This makes it unique, because it is not prone to choice of dataset. Every inference learns the prior from given spectrogram from noise.

III. Experiments

A. Dataset and preprocessing

For evaluation, we use 10 music signals from [16] with different harmonic and percussive characteristics (violin, accordion, celesta, etc.). The signals are resampled from the original 44.1 kHz to 16 kHz and shortened to 5 seconds. This step was necessary for the use of a pre-built network.

B. Time-frequency masks

We tested two types of TF masks. In both cases, we define the ratio of missing coefficients expressed as a percentage. We are aware of the rounding error due to the fact that some percentage ratios of total number of coefficients ($M \cdot N$) are floats. However, the error does not exceed tenths of a percent and we neglect this small mistake. The range of absent percentages extends from 10 to $90\,\%$ in increments of ten, comprising a total of ten masks. Masks are applied to same complex numbers (2D mask for real part is duplicated for the imaginary part of complex coefficients).

The first variant chooses the specified ratio of random positions from the whole coefficient matrix. To be fair to the whole dataset, we use a single pregenerated mask for all ten signals in the dataset. This ensures that coefficients on same positions are vanished.

The second mask is generated for each signal independently since it depends on the context of the signal. We calculate the absolute value of every complex coefficient (i.e. energy of coefficients). Then we sort the coefficients in ascending order by energy separately for every column. We remove the expected ratio of coefficients in each column.

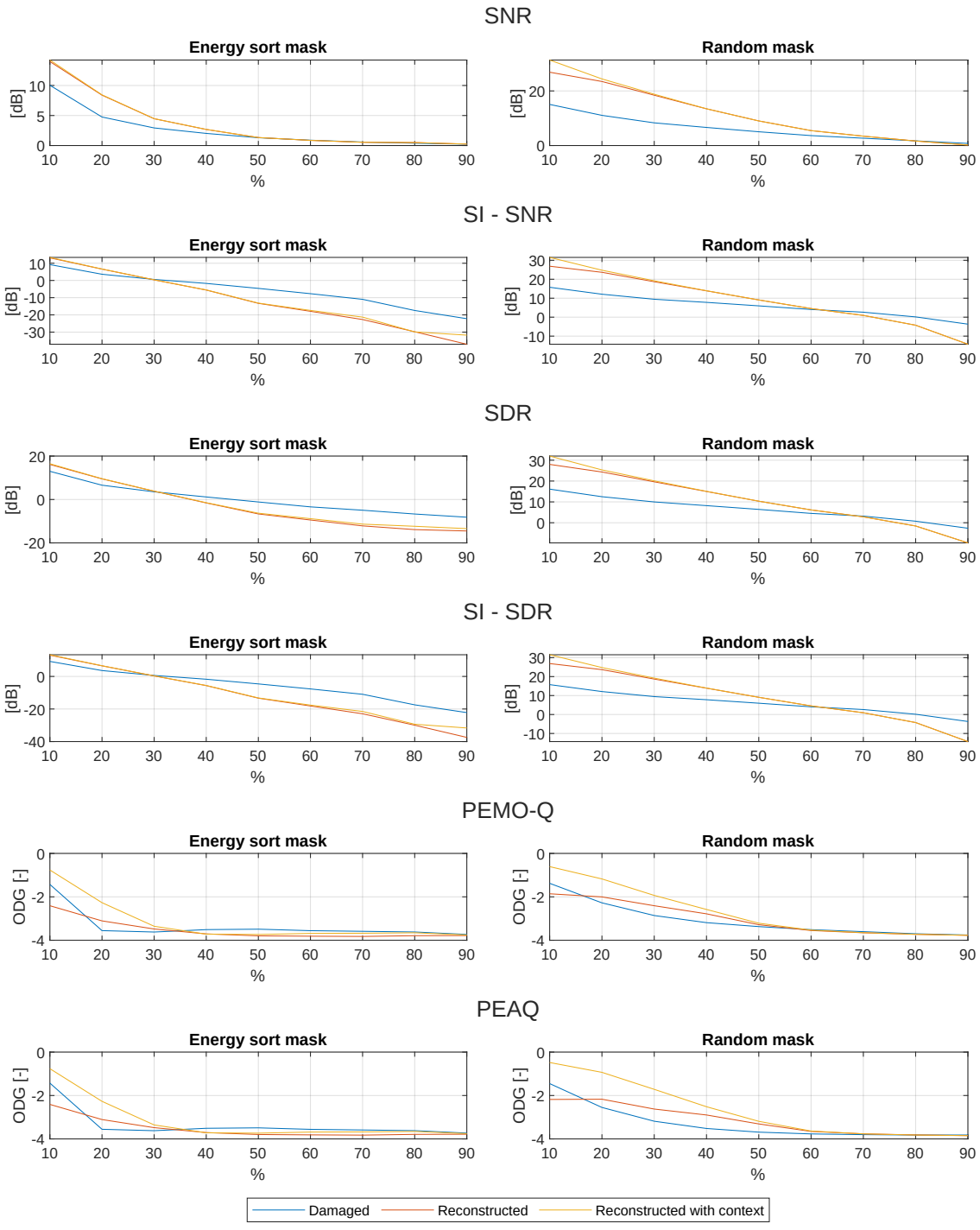

Fig. 1. The performance of the metrics used for both masks shows that the random mask outperforms the energy significance-based sorting mask.

C. Network architecture

The very trending neural architecture in the area of audio processing is UNet [17]. It has a very similar structure to codec itself, because it consists of two principal parts, encoder and decoder. The main building blocks of these networks are convolutional layers, pooling layers, and skip connections.

In particular, the implementation of DPAI [12] uses the MultiResUNet [18] architecture with harmonic convolutions [19]. This operation speeds up optimisation and has better results from the perceptual point of view. A music signal has a nonlinear frequency representation. This makes the adjacent coefficients and 2D convolution quite inappropriate. This problem is solved by what the authors call the harmonic convolution, which computes its value from multiples of chosen coefficient rather than neighbouring. These operations should be more suitable for an audio signal because the harmonic series is composed of integer multiples of the base frequency [20].

We used Miotello's implementation[1] where we modified only the TF masks. The net specifications described as "best2" were used as in [12]. This model has approximately 7 million parameters. The complex spectrogram is split into real and imaginary parts and stacked in the network. The simplified network architecture is shown in figure 2.

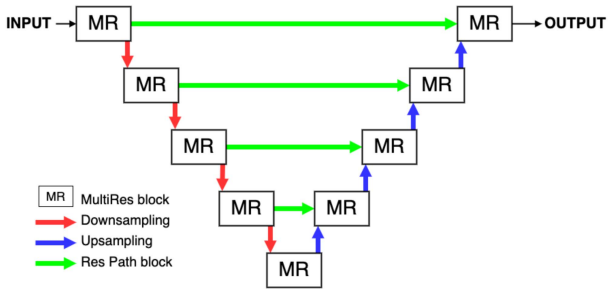

Fig. 2. MultiResUNet architecture used from [12].

D. Metrics

We only use objective metrics for the evaluation. From non-perceptive methods, signals were evaluated by signal-to-noise ratio (SNR) and its updated version Scale-Invariant SNR (SI-SNR) and also signal-to-distortion ratio (SDR) and its scale-invariant version (SI-SDR). In [21] compression was described as one of the scenarios in which classic SNR/SDR results should be evaluated carefully and instead use scale-invariant versions. As the perceptually motivated metrics, PEMO-Q [22] and PEAQ [23] were used. The evaluated signals for the last two metrics were resampled to 48 kHz and 44.1 kHz, respectively, due to its implementation limitations. For every metric, its higher value means a better reconstruction.

IV. Results and Discussion

A single reconstructed signal was generated approximately in 20 minutes. Each mask has two possible outputs. The first is the output generated without any post-processing. Second, called "with context", replace generated samples after inference with reliable coefficients before inference (where mask has true value).

In figure 1, the metrics evaluated for both masks are displayed. The variant that uses a random mask outperformed the other in all metrics discussed. These outcomes are similar to audible results. Specifically, for the random mask, satisfactory results are obtained when approximately 50 to 60 % of the coefficients are missing. In contrast, for the more meaningful mask with coefficients sorted by energy, the results are regrettably inferior, with noticeable changes occurring when 20 to 30 % of coefficients in the spectrogram are absent.

V. Conclusion

As the first in research community, we present the redesign of the deep-audio prior audio inpainting framework for the task of audio compression. The results have shown that the concept different from inpainting works and leads to a reasonable solution. However, application in the real world is difficult. Mainly due to a very long inference time, which is very hard to reduce in prior based methods.

The success of random initialisation offers new ways to explore new ideas. One idea is to implement deep prior for the raw waveform, but inference time could be similar. Today, the most trending model is the diffusion model, which also recovers information from noisy distributions and could be tested in further research.

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

References

[1] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[2] Y. Ai, X.-H. Jiang, Y.-X. Lu, H.-P. Du, and Z.-H. Ling, "Apcodec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding," *arXiv preprint arXiv:2402.10533*, 2024.

[3] D. Pan, "A tutorial on mpeg/audio compression," *IEEE MultiMedia*, vol. 2, no. 2, pp. 60–74, 1995.

[4] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, "Iso/iec mpeg-2 advanced audio coding," *J. Audio Eng. Soc*, vol. 45, no. 10, pp. 789–814, 1997. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10271

[5] J. Moffitt, "Ogg vorbis—open, free audio—set your media free," *Linux journal*, vol. 2001, no. 81es, pp. 9–es, 2001.

[6] P. Peter, J. Contelly, and J. Weickert, "Compressing audio signals with inpainting-based sparsification," in *Scale Space and Variational Methods in Computer Vision*, J. Lellmann, M. Burger, and J. Modersitzki, Eds. Cham: Springer International Publishing, 2019, pp. 92–103.

[7] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2022.

[8] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[9] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[10] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," *International Journal of Computer Vision*, vol. 128, no. 12, 2020. [Online]. Available: https://doi.org/10.1007/s11263-020-01303-4

[11] Y. Tian, C. Xu, and D. Li, "Deep audio prior," *arXiv preprint arXiv:1912.10292*, 2019.

[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 Transactions on Audio, Speech, and Language Processing*, pp. 1–11, 2023.

[13] 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, Sep. 2020-21. [Online]. Available: https://dafx2020.mdw.ac.at/proceedings/index.html

[14] A. Marafioti, P. Majdak, N. Holighaus, and N. 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.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[16] (2008) EBU SQAM CD: Sound quality assessment material recordings for subjective tests. online. [Online]. Available: https://tech.ebu.ch/publications/sqamcd

[17] A. A. Nair and K. Koishida, "Cascaded time + time-frequency UNet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 7153–7157.

[18] N. Ibtehaz and M. S. Rahman, "Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608019302503

[19] H. Takeuchi, K. Kashino, Y. Ohishi, and H. Saruwatari, "Harmonic lowering for accelerating harmonic convolution for audio signals," 10 2020, pp. 185–189.

[20] P. Balazs, N. Holighaus, T. Necciari, and D. Stoeva, *Frame Theory for Signal Processing in Psychoacoustics*. Cham: Springer International Publishing, 2017, pp. 225–268. [Online]. Available: https://doi.org/10.1007/978-3-319-54711-4_10

[21] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[22] 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.

[23] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," MMSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University, Tech. Rep., May 2002.