Audio Engineering Society

Conference Paper

Presented at the International Conference on
Audio Archiving, Preservation & Restoration
2023 June 1–3, Culpeper, VA, USA

Noise reduction in analog tape audio recordings with deep learning models

Ignacio Irigaray[1], Martin Rocamora[1], and Luiz W. P. Biscainho[2]

[1] *Universidad de la República, Uruguay*
[2] *Universidad Federal do Rio de Janeiro, Brazil*

Correspondence should be addressed to Ignacio Irigaray (irigaray@fing.edu.uy)

ABSTRACT

This work addresses the problem of noise reduction in tape recordings using a deep-learning approach. First, we build a data set of audio snippets of tape noise extracted from different functional tape equipment — comprising open reel and cassette. Then, we adapt and train an existing deep-learning architecture originally proposed to remove noise from 78 RPM gramophone records. The model learns from mixtures of the noise snippets with clean audio excerpts at different SNRs. Experimental results validate the approach, showing the benefits of using real tape recording noise in training the model. Furthermore, the data set of tape noise snippets and the trained deep-learning models are publicly available. In this way, we encourage the collective improvement of the data set and the broad application of the denoising approach by sound archives.

1 Introduction

Magnetic tape recording was the dominant technology for audio recording during several decades of the XX century. It has been widely available since the 1950s and was gradually taken over by digital audio recording technology since the standardization of the compact disc format in 1980. However, magnetic audio tape formats are now obsolete, so the only way to preserve tape recordings and make them accessible is their digitization and transfer to safe digital repositories as long as replay equipment is in operable condition and the tapes have not deteriorated [1, 2, 3]. After the sound transfer process, either from tapes or discs, digital audio restoration takes place to treat different types of disturbances and degradation, such as thumps, clicks, and hiss [4].

Traditional methods for digital audio restoration are based on Digital Signal Processing (DSP) techniques, such as Wiener filtering and autoregressive (AR) modelling [5, 6, 7]. However, the significant progress brought by deep learning [8] to computer vision [9] and natural language processing [10] has also extended to the audio domain, improving the state-of-the-art in problems like speech recognition [11] and sound source separation [12]. Consequently, some recent works have addressed audio restoration tasks using a deep learning approach [13], including audio upsampling [14], bandwidth extension [15], and denoising [16].

The work by Moliner et al. [17] is particularly relevant to the present paper since it proposes a U-Net model for noise reduction inspired by [16] and its application to 78 RPM gramophone recordings. The method can suppress colored noise, rumble, and impulsive events [17]. This versatility is a clear advantage compared to the DSP approach, in which different techniques must be applied to different types of imperfection. For instance, while stationary noise is typically treated by spectral subtraction [18], clicks and thumps are treated independently by firstly detecting them and then interpolating the missing samples [17, 19]. According to the authors, one of the keys to the high performance reported in [17] was the use of more realistic noise data compared to their inspiring work [16], which uses a similar deep network architecture. The construction of the noise data set was possible thanks to a collaborative project to massively digitize 78 RPM records.[1]

In this work, we draw upon Moliner et al. [17] and conduct a series of experiments to test the validity of their deep-learning approach for denoising tape recordings. To that end, we build a data set of audio snippets of tape noise extracted from different functional tape equipment — comprising open reel and cassette. We then train a deep learning model using the architecture proposed in [17] on mixtures of the noise snippets with clean audio excerpts at different levels of signal-to-noise-ratio (SNR). Finally, we evaluate the obtained model on a test data set using objective methods and compare it to the model released in [17] and to a traditional noise reduction method [18]. The results attest the effectiveness of the approach for noise reduction in analog tape audio recordings, showing the benefits of using real tape recording noise for training the model.

The data set of tape noise audio fragments and the trained deep-learning models are being released for public access with the publication of this paper. We encourage the collective improvement and expansion of the data set by contributions of individuals and institutions. We believe that sound archives can play a key role in extending the noise data set with samples of their operational equipment. An improved data set would allow for training better models, either more general or targeted to specific tape recording devices.

The rest of the paper is organized as follows. The next section describes the deep-learning model, the training strategy, and the noise and clean-audio data sets used for training and evaluation. Section 3 presents the experiments and results. The paper ends with a critical discussion and some directions for future work.

2 Method

2.1 Clean and noisy data collection

Two data sets are used to train the model and evaluate its performance: one with clean music recordings and another with audio fragments of tape noise. The two data sets are artificially combined to simulate the effect of real tape recordings by adding tape noise to the clean music audio. The process is represented by Equation 1, where y, and z correspond to clean music and tape noise audio fragments, respectively, whilst x is the simulated tape audio recording segment. The α parameter controls the SNR, while the β parameter controls the scale factor. As a data augmentation scheme the SNR and β are chosen from a log-uniform distribution between 6 and 32 dB, and from 0 to -6 dB, respectively, for the training stage.

$$x = \beta(y + \alpha z) \tag{1}$$

Each data set was split into train, validation, and test, as described in the following.

2.1.1 Data set of clean music audio

As in [17], the clean music audio is taken from the MusicNet[2] data set [20]. It is a collection of 330 freely — licensed classical music recordings for a total duration of 34 hours, commonly used for training models and as a benchmark for comparing results. As available, the data is organized according to the train/test split described and used in [21], in which only a small (1%) but representative subset is selected for testing. We stuck to the test set but divided the remaining data into 10% for validation and 90% for training.

2.1.2 Data set of tape noise audio

For building the analog tape noise data set, blank tapes were reproduced in different replay devices, and their output was digitized using an M-Audio Fast Track Pro audio interface at 44.1 kHz sampling rate. The equipment used was available at the National Center for Music Documentation[3], where the first author acts as a technical consultant. All the devices were serviced and calibrated before doing the tape noise recordings.

Five open reel replay devices were used: two semi-professional Revox A77 recorders — one normal-speed

[1] The Great 78 Project: https://great78.archive.org/

[2] Avaiable from zenodo: https://doi.org/10.5281/zenodo.5120004

[3] http://www.cdm.gub.uy/ (Montevideo, Uruguay)

model (NS) and one high-speed model (HS) —, a vintage tube Revox C-36 recorder, and two portable Uher recorders — a 4000 report S and a 4000 report L. The blank 1/4" open reel tape used was a Premium Analog Recording Tape by ATR Magnetics. In addition, a double deck Technics TR-575 compact cassette player with a blank TDK-HX-S60 cassette was recorded at nominal speed. Each of the cassette decks was recorded separately. Table 1 summarizes all the replay devices used and at which speed they were recorded. An illustrative picture of each device is shown in Figure 1.

Table 1: Replay devices used for building the tape noise data set and their corresponding speeds.

Recorder	Speeds (IPS)
Revox A-77 (NS)	7.5 3.75
Revox A-77 (HS)	7.5 15
Revox C-36	7.5 3.75
Uher 4000 L	3.75 1.875
Uher 4000 S	3.75 1.875
Technics TR-575	1.875

The tape noise data set totals 2 hours of audio, corresponding to 10 minutes for each device and speed combination (considering the two cassette decks), and it is released for public access with this publication.[4]

2.2 Noise reduction model

2.2.1 Network architecture

A two-stage U-net [22] architecture with a supervised attention module (SAM) initially proposed in [17] was utilized for music denoising. The input of both stages is the complex-valued short-time Fourier transform (STFT) of the noisy signal—treated as two real-valued separate channels—appended with a frequency-positional embedding. Let x be the audio fragment of the noisy input signal sampled at 44.1 kHz. The STFT of x, denoted as X, is computed with a window length of $N = 2048$ and a hop size of $h = 512$ samples.

The first stage estimates the STFT of the residual noise fragment, $\hat{Z}$, and passes it to the SAM to propagate only the relevant features to the second stage. Additionally, the STFTs of the residual noise estimate and the noisy input signal are summed to get an estimation of the STFT of the clean signal, i.e. $\hat{Y}_1 = X + \hat{Z}$.

The STFT of the clean signal is also estimated in the second stage, denoted as $\hat{Y}_2$, from the output of the first stage and the STFT of the noisy input signal. According to the authors of [17], this two-stage schema minimizes the occurrence of annoying musical noise artifacts. We refer the reader to [17] for further details of the network architecture.

2.2.2 Training process

The loss function of equation 2 is minimized during training. The mean absolute error between the output of both stages (i.e. the estimates of the STFT of the clean signal, Y_1^k and Y_2^k) and the STFT of the clean signal Y^k is calculated for each bin k of the STFT.

$$\mathscr{L} = \frac{1}{K}\sum_k (|\hat{Y_1}^k - Y^k| + |\hat{Y_2}^k - Y^k|). \tag{2}$$

The Adam optimizer was used with parameters $\beta_1 = 0.5$, $\beta_2 = 0.9$, the learning rate was initialized in $1e-4$ and divided by 10 every 100.000 steps. The model was trained for 320000 steps in an RTX 3090 GPU with 24 Gb of RAM, and the training time was 48 hours. The value of the loss for the train set and the validation set were used for the training stopping criteria.

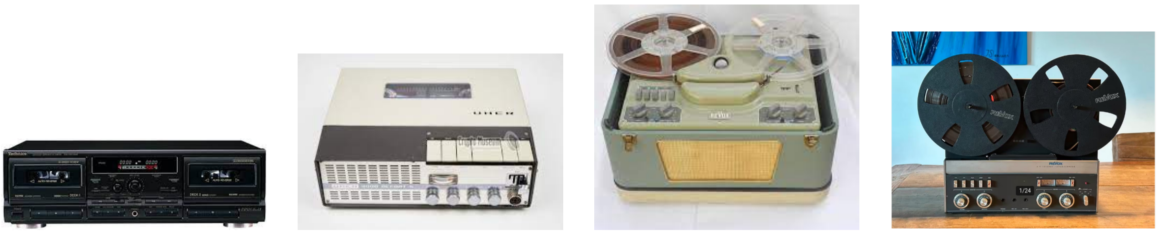
3 Experiments and results

The performance of the model trained on the tape noise data (Tape Noise Model) was assessed on a test data set using objective evaluation measures. For comparison, two other methods were also tested with the same evaluation setup: the model released with the original paper [17] — which was trained on 78 RPM noise data (78 RPM Noise Model) —, and the traditional spectral subtraction method (Spectral Subtraction) [18] — using the `noisereduce`[5] python package [23].

The test data set was build combining clean music fragments with tape noise fragments taken from the corresponding test sets. A total of 100 fragment pairs were selected, with a duration of 10 seconds. The fragments were combined at two different SNRs: 10

[4] `https://github.com/IgnacioIrigaray/AnalogAudioTapeDenoising`

[5] `https://pypi.org/project/noisereduce/`



Technics TR-575 Uher 4000 report L Revox C36 Reovx A77

Fig. 1: Type of replay devices used in the tape noise data set.

Table 2: Results of the evaluation experiment.

SNR	Method	Δ SNR	Δ PEAQ
10dB	Tape Noise Model	8.32	1.74
	78 RPM Noise Model	4.23	0.82
	Spectral Subtraction	-7.31	0.23
16dB	Tape Noise Model	4.61	1.76
	78 RPM Noise Model	1.89	1.16
	Spectral Subtraction	-13.2	0.26

dB simulating an adverse recording scenario, and 16 dB for a more favorable one. Note that in the experiments in [17] an SNR of 3 dB is also reported, which may be reasonable for 78 RPM recordings but is not realistic for tape recordings and was therefore discarded.

Two objective evaluation metrics were computed for each fragment comparing the original noisy music signal with the denoised output signal: the *SNR* and the perceptual metric *PEAQ* [24]. The `gstreamer` [25] plugin `gstPEAQ` [26] was used for computing the PEAQ evaluation measure. The average gain Δ SNR and Δ PEAQ over the whole test set is reported in Table 2 for each method and for the two SNR conditions.

4 Discussion

In this work we applied a deep-learning approach to the problem of noise reduction in tape recordings. To do that, we build a data set of audio excerpts of tape noise by reproducing blank tapes in different replay devices and digitally recording their output. The tape noise data set is then combined with clean music recordings to train an existing deep-learning architecture originally proposed to remove noise from 78 RPM gramophone records [17]. Finally, we evaluate the obtained model on a test data set using the SNR and PEAQ objective measures, and compare it to the model released in [17] and to a traditional noise reduction method [18]. The experimental results, see Table 2, show the effectiveness of the deep learning approach producing a noise reduction greatly superior to the one of the traditional method. Besides, the evaluation also shows the benefits of using real tape recording noise for training the model, since the model trained on tape noise samples clearly outperforms the model trained on noise extracted from 78 RPM discs.

The data set of tape noise fragments and the trained deep-learning models are being released for public access with the publication of this paper.[6] We plan to extend the number of replay devices on the data set and we welcome external contributions of individuals and institutions. We believe that sound archives can play a key role in extending the noise data set with samples of their operational equipment.

There are several possible strands for future work. Instead of a generic noise reduction model like the one produced in this work, over-fitting could be beneficial in some application scenarios. In future works, we plan to apply fine-tuning techniques to produce models for denoising adapted to a specific replay device. In addition, subjective evaluation tests should be conducted to perceptually validate the results obtained in this work.

5 Acknowledgments

We thank the National Center for Music Documentation (Montevideo, Uruguay) for their support.

[6] `https://github.com/IgnacioIrigaray/AnalogAudioTapeDenoising`

References

[1] Committee, I. T. et al., "The IASA-TC 03: The Safeguarding of the Audiovisual Heritage: Ethics, Principles and Preservation Strategy [online].[cited 2018.2. 17]," 2017.

[2] Tovell, A., "Audio Preservation and Access: Overecoming the Challenges," in *Audio Engineering Society Conference: 2018 AES International Conference on Audio Archiving, Preservation & Restoration*, Audio Engineering Society, 2018.

[3] Pace, A., "Magnetic Tape Alert Project report," Technical report, International Association of Sound and Audiovisual Archives, 2020.

[4] Esquef, P. A., *Audio Restoration*, pp. 773–784, Springer New York, New York, NY, 2008, ISBN 978-0-387-30441-0, doi:10.1007/978-0-387-30441-0_40.

[5] Godsill, S. J. and Rayner, P. J. W., *Digital Audio Restoration - a statistical model based approach*, Springer-Verlag, Berlin, Heidelberg, 1998.

[6] Dewasthale, M. M. and Kharadkar, R., "Acoustic noise cancellation using adaptive filters: A survey," in *2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies*, pp. 12–16, IEEE, 2014.

[7] Yu, G., Mallat, S., and Bacry, E., "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal processing*, 56(5), pp. 1830–1839, 2008.

[8] Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.

[9] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," in F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, Curran Associates, Inc., 2012.

[10] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., "Deep Contextualized Word Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, Association for Computational Linguistics, New Orleans, LA, USA, 2018, doi:10.18653/v1/N18-1202.

[11] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, 29(6), pp. 82–97, 2012, doi:10.1109/MSP.2012.2205597.

[12] Aditya Arie Nugraha, A. L. and Vincent, E., "Multichannel Audio Source Separation With Deep Neural Networks," in *IEEE Transactions on audio, speech, and language processing, Vol. 24, No. 9*, pp. 1652–1664, 2016.

[13] Lee, J. and Han, S., "NU-Wave: A Diffusion Probabilistic Model for Neural Audio Upsampling," in *Proc. Interspeech 2021*, pp. 1634–1638, 2021, doi:10.21437/Interspeech.2021-36.

[14] Deng, J., Schuller, B., Eyben, F., Schuller, D., Zhang, Z., Francois, H., and Oh, E., "Exploiting time-frequency patterns with LSTM-RNNs for low-bitrate audio restoration," *Neural Computing and Applications*, 32(4), pp. 1095–1107, 2020, doi:10.1007/S00521-019-04158-0.

[15] Moliner, E. and Välimäki, V., "BEHM-GAN: Bandwidth Extension of Historical Music using Generative Adversarial Networks," *arXiv preprint arXiv:2204.06478*, 2022.

[16] Li, Y., Tagliasacchi, M., Gfeller, B., and Roblek, D., "Learning to Denoise Historical Music," in J. Cumming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, and T. de Reuse, editors, *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, pp. 504–511, 2020.

[17] Moliner, E. and Välimäki, V., "A Two-stage U-Net for high-fidelity denoising of historical recordings," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), pp. 841–845, IEEE, 2022.

[18] Boll, S., “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on acoustics, speech, and signal processing*, 27(2), pp. 113–120, 1979.

[19] de Carvalho, H. T., Avila, F. R., and Biscainho, L. W. P., “Bayesian Restoration of Audio Degraded by Low-Frequency Pulses Modeled via Gaussian Process,” *IEEE Journal of Selected Topics in Signal Processing*, 15(1), pp. 90–103, 2021, doi:10.1109/JSTSP.2020.3033410.

[20] Thickstun, J., Harchaoui, Z., and Kakade, S. M., “Learning Features of Music from Scratch,” in *International Conference on Learning Representations (ICLR)*, 2017.

[21] Thickstun, J., Harchaoui, Z., Foster, D. P., and Kakade, S. M., “Invariances and data augmentation for supervised music transcription,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2241–2245, IEEE, 2018.

[22] Ronneberger, O., Fischer, P., and Brox, T., “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[23] Sainburg, T., Thielk, M., and Gentner, T. Q., “Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires,” *PLoS computational biology*, 16(10), p. e1008228, 2020.

[24] Thiede, T., “PEAQ–The ITU Standardfor ObjectiveMeasuremenot f Perceived Audio Quality,” *J. Audio Eng. Soc.*, 48(1), p. 27, 2000.

[25] Newmarch, J. and Newmarch, J., “GStreamer,” *Linux Sound Programming*, pp. 211–221, 2017.

[26] Holters, M. and Zã, U., “GstPEAQ – an Open Source Implementation of the PEAQ Algorithm,” p. 4, 2015.