# MagTapeDB: A Dataset of Historical Magnetic Tape Recordings

Ignacio Irigaray
*Instituto de Ingeniería Eléctrica*
*Universidad de la República*
Montevideo, Uruguay
irigaray@fing.edu.uy

Emilio Martinez
*Instituto de Ingeniería Eléctrica*
*Universidad de la República*
Montevideo, Uruguay
emartinez@fing.edu.uy

Diego Silvera Coeff
*Instituto de Ingeniería Eléctrica*
*Universidad de la República*
Montevideo, Uruguay
dsilveracoeff@fing.edu.uy

Luiz W. P. Biscainho
*DEL/Poli & PEE/COPPE*
*Universidad Federal do Rio de Janeiro, Brazil*
Rio de Janeiro, Brazil
wagner@smt.ufrj.br

***Abstract*—We present a novel dataset designed to support the development and evaluation of audio restoration techniques focused on historical music recordings. The dataset comprises over 800 audio excerpts, including musicological tape recordings, isolated tape hiss segments, and pitchpipe tones used as tuning references. Each fragment is annotated with metadata such as instrument presence, year of recording, and tape reel number. The dataset enables a variety of restoration-related tasks, including denoising, noise profiling, instrument detection, and segmentation. We also provide baseline results for denoising using state-of-the-art deep learning models and demonstrate an application for playback speed correction using Electrical Network Frequency (ENF) analysis. Our goal is to contribute to the preservation of audio heritage by facilitating reproducible research and benchmarking in music restoration.**

***Index Terms*—audio restoration, magnetic tape, cultural heritage, deep learning, denoising, Electrical Network Frequency, dataset**

## I. Introduction

Magnetic tape was one of the most widely used media for recording music during the second half of the 20th century. From professional multitrack studio sessions to field recordings and home archives, a significant portion of the world's sonic heritage is stored on this medium. However, tape-based recordings are inherently fragile and prone to various forms of degradation over time, including hiss, hum, dropouts, wow and flutter, and saturation. These artifacts can severely impact the intelligibility and aesthetic value of the recordings, posing a major challenge for both restoration and long-term preservation.

In recent years, deep learning techniques have shown great potential in audio restoration tasks such as denoising, source separation, and inpainting. However, the development and evaluation of these methods frequently depend on synthetic or mismatched datasets that do not accurately reflect the specific characteristics of analog tape degradation. Although large-scale datasets exist for tasks like speech enhancement and music source separation, there remains a significant lack of publicly available resources that capture the distinctive noise profiles and distortions found in magnetic tape —— particularly those originating from musicological fieldwork.

To address this gap, we introduce **MagTapeDB**, a curated dataset composed of over 800 audio clips.[1] The dataset includes music fragments affected by analog tape degradation, isolated tape noise excerpts, and short pitchpipe recordings used as tuning or speed references. The collection features field recordings, music with varied instrumentation, and some interviews. Each recording is accompanied by metadata that includes information such as instrumentation, year of recording, tape reel number, etc. The dataset is specifically focused on archival and traditional music, genres that are typically underrepresented in contemporary machine learning datasets. This specialization reflects the origin and purpose of the collection, which is grounded in musicological field recordings and heritage preservation efforts.

To illustrate the potential uses of the dataset, we present baseline results for two restoration tasks: (1) recording speed correction, using both Electrical Network Frequency (ENF) analysis and pitchpipe tone alignment, which also enables the estimation of the pitchpipe's nominal tuning frequency; and (2) denoising, implemented using a U-Net-based deep learning architecture.

Although the current release of **MagTapeDB** includes only 27 of the 127 available tapes, it represents the first stage of a long-term digitization and preservation effort carried out in collaboration with the Lauro Ayestarán Center for Musical Documentation. While this initial scale may limit the direct training of large data-hungry deep learning models, the ongoing expansion of the digitized collection is designed to

[1]The dataset, accompanying code and examples are available at https://github.com/IgnacioIrigaray/MagTapeDB/

address this limitation by progressively increasing the dataset's volume and diversity.

**MagTapeDB** is intended as a resource for developing tools that support the preservation and valorization of cultural audio heritage.

## II. Related Work

Traditional methods for digital audio restoration are based on Digital Signal Processing (DSP) techniques, such as Wiener filtering and autoregressive (AR) modelling [1]–[3]. In the last decade, however, deep learning approaches have emerged as powerful alternatives, offering improved performance and generalization across diverse tasks such a in audio processing tasks, including speech recognition [4] and sound source separation [5].

Several large-scale datasets have played a crucial role in the development of these methods. In the speech domain, corpora like VoiceBank+DEMAND [6], DNS Challenge datasets [7], and WHAMVox dataset [8] provide paired clean/noisy samples and have enabled standardized evaluation of denoising algorithms. In music, datasets such as MUSDB18 [9], Slakh2100 [10], and MedleyDB [11] have supported research in source separation and music analysis.

Denoising often relies on supervised machine learning models, which require paired examples of clean and degraded audio for training [12], [13]. In these works, real gramophone or magnetic tape noise is artificially added to clean audio signals in order to generate such training pairs.

In contrast, degradations from magnetic tape include a wide range of artifacts: high-frequency hiss from bias oscillators, power-line hum, low-frequency instability (wow and flutter), saturation distortion, and physical defects such as dropouts or tape crinkle. These characteristics are rarely present in existing benchmarks, making it difficult to train or evaluate restoration methods for archival materials.

A few studies have addressed analog degradation modeling, including work on tape emulation for audio effects, and recent applications of neural networks to simulate analog distortion. However, datasets used in those works are often small, not publicly available, or focused on specific use cases (e.g., guitar amp modeling, or tape delay emulation).

The field of audio pattern recognition —— broadly encompassing classification, detection, and segmentation tasks —— has seen substantial advances through the use of convolutional neural networks and, more recently, transformer-based models. However, the application of these techniques to degraded archival recordings remains underexplored, in part due to the lack of representative training data.

**MagTapeDB** aims to bridge this gap by offering a curated dataset of real or simulated tape degradations applied to music recordings. It is designed to support both restoration-focused research and more general pattern recognition tasks, such as the classification of degradation types, temporal segmentation of noise events, or contrastive learning of clean vs. degraded audio.

By introducing a standardized, open dataset that captures the complexities of analog noise, **MagTapeDB** contributes to ongoing efforts to make machine learning methods more applicable to cultural heritage preservation.

## III. Dataset Description

Between 1943 and 1966, musicologist Lauro Ayestarán (Montevideo, 1913 – Montevideo, 1966) dedicated himself to the task of collecting popular and traditional music throughout Uruguay, traveling the country with sound recording equipment (Figure 1). This was a unique endeavor in terms of its scope, resulting in a legacy of over 3000 sound recordings, detailed field notes, lyric transcriptions and melodic notations, photographic negatives and prints, numerous articles, a seminal book —— the first volume of La Música en el Uruguay —— as well as statistical tables, thematic folders, sketches, maps, and catalogs. Taken as a whole, this body of work constitutes a singular documentary resource that provides essential tools for understanding Uruguay's musical identity.

Fig. 1: Lauro Ayestarán recording Ramón López. Aguas Corrientes, Canelones Department. January 14, 1962. Photo by Juan Carlos Santurión. CDM Archive. Note the use of a battery-powered recorder.

The recordings captured a wide range of musical expressions, including criolla songs and acriollada dances, candombe, murga, children's songs, tango, northern folk repertoire, street vendor chants, and the singing of cart drivers. The digitization of the complete collection of recordings is currently underway and will be available in future releases of this dataset. This initial version of the dataset includes the digitized content of 27 tapes out of a total of 127, a sample of which is shown in Figure 2. Table I provides a breakdown of the 894 audio excerpts included in this initial release. The majority correspond to musical recordings, followed by isolated noise segments and pitchpipe tones.

While musical recordings represent the most numerous and diverse content, noise excerpts account for a significant portion of the total duration due to the inclusion of long background

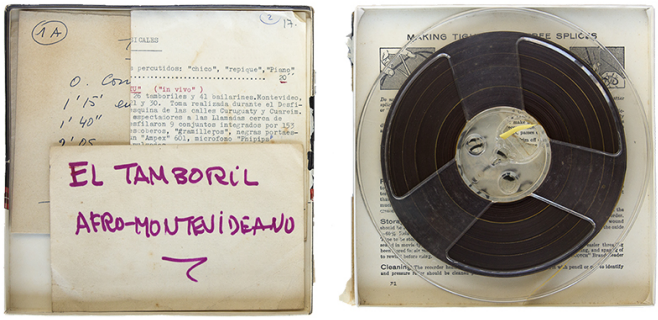


Fig. 2: Magnetic tape reel and original case, illustrating the type of archival material used as the source for **MagTapeDB**.

TABLE I: Summary of digitized tapes in the current release.

| Category | Number of recordings | Duration (min) |
|---|---|---|
| Musical recordings | 321 | 161.5 |
| Pitchpipe tones | 276 | 3.42 |
| Noise excerpts | 297 | 186 |
| **Total** | **894** | **350.92** |

segments. Pitchpipe tones, although short in duration, appear in approximately 86% of the recordings and play a key role in playback speed estimation and tuning analysis.

The tapes are 1/4-inch mono half-track recordings, originally recorded at speeds of 3.75 and 7.5 inches per second (ips). Playback was performed using a Revox A77 tape recorder, which underwent a full service, including the replacement of the playback head with a new unit. The audio interface used was a Universal Audio Apollo Solo.

Gain levels were adjusted to ensure that peak values remained below –6 dBFS. Each tape was played back in both directions, and both channels were recorded, since audible differences have been reported when playing tapes in reverse, as discussed in [14].

In addition to the audio files, the dataset includes a CSV file containing detailed metadata for each recording. The fields provided are: recording ID, genre or category (especie), title, instruments, recording location (site and locality), date, tape speed, and the estimated tuning fork frequency when available. This structured metadata enables researchers to filter the dataset by content type, geographic origin, instrumentation, or technical attributes relevant to restoration and analysis tasks.

The digitization process is illustrated in Fig. 3. The analog tape is played back using a Revox reel-to-reel recorder, captured through a Universal Audio interface, and subsequently digitized, segmented, and stored. The digitization is performed using the Reaper software, followed by a manual annotation workflow carried out in three stages. First, each region of interest in the audio is labeled with its corresponding recording ID. Second, the pitchpipe tone —— when present —— is annotated. Finally, segments containing only background noise are identified. Based on these labeled regions, the individual audio files are extracted, those shorter than 30 s are discarded, and a centralized 30-second excerpt is selected from each to form the final dataset entries.

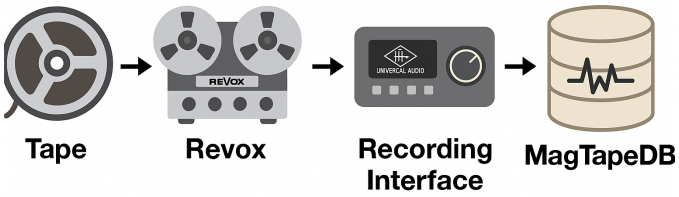


Fig. 3: Signal digitization workflow for **MagTapeDB**. The analog tape is played back using a Revox reel-to-reel recorder, captured via a Universal Audio interface, and subsequently digitized, segmented, and stored in the **MagTapeDB** collection

**MagTapeDB** is freely available for non-commercial scientific research. The dataset, along with metadata and baseline code, can be accessed through a public repository[2]. By making this resource openly available, we aim to foster reproducible research, encourage collaboration across disciplines, and support the development of new methods for the analysis and restoration of degraded analog recordings.

## IV. Baseline Experiments

### A. Speed Correction using ENF

Accurate playback speed is essential for the faithful perception and analysis of archival recordings. Ideally, the playback speed should match the original recording speed as closely as possible. However, several factors may lead to speed deviations, including uncalibrated equipment, unstable power supplies (which were common in remote locations at the time of the recordings or when using battery-powered devices), and tape degradation.

Early musicologists were already aware of the importance of playback speed, and in many cases included a tuning fork or pitch pipe at the beginning of the recording as a reference tone —— a practice found in a significant number of items in this dataset. While these reference tones are extremely valuable for playback speed estimation, they present two main limitations. First, not all recordings include such a reference. Second, the nominal frequency of the pitch pipe is often unknown. Since international pitch standardization was not fully established until 1955[3], we cannot assume that the reference tone corresponds to A440. Additionally, the tuning frequency may have changed over time due to environmental factors or material degradation. For instance, wooden pitch pipes are susceptible to shrinkage, which can lead to an increase in pitch [16].

The Electrical Network Frequency (ENF) corresponds to the standard frequency of the AC power grid, usually 50 or 60 Hz depending on the region. During a magnetic tape

[2]https://github.com/usuario/magtapedb

[3]Although the frequency of 440 Hz for the note $A_4$ had already been proposed at the Stuttgart Conference in 1834, it was not formally adopted as an international standard until 1955, when the Acoustical Committee of the International Organization for Standardization met in London and reaffirmed the recommendation of 440 Hz for tuning [15]. A440 has remained the official international reference pitch since then.

recording, electromagnetic interference from nearby electrical sources can leave traces of this frequency in the audio signal, typically in the form of low-amplitude hums or modulations at the fundamental ENF (e.g., 50 Hz) and its harmonics. One common application of ENF analysis is in forensic audio, where it is used to verify authenticity and detect tampering in recordings [17]–[19].

When the tape's recording or playback speed deviates from the original, the frequency of the embedded ENF shifts accordingly. Deviations smaller than 50 millihertz ($\Delta f < 50\,\mathrm{mHz}$) are typically considered to fall within the normal operating range of the power grid [19]. Although tolerances may have been slightly broader in the mid-20th century, ENF during that period can still be regarded as sufficiently stable for analysis and correction purposes. Since its fluctuations are recorded along with the audio material on tape, estimating the embedded ENF signal enables the correction of playback speed deviations.

In this work, we use the correlation between the frequency of the pitchpipe tone and the estimated ENF component to infer the most likely nominal frequency of the pitchpipe. This information allows for more accurate playback speed correction in cases where pitchpipe tones are present but ENF traces are absent —– typically when the recording was made using battery-powered equipment.

To estimate the fundamental frequency of the pitchpipe tones, we use the probabilistic YIN algorithm (pYIN) [20] implemented in the librosa library. This method extends the original YIN algorithm by modeling pitch candidates probabilistically and applying a Hidden Markov Model (HMM) to track pitch over time. It combines temporal autocorrelation with statistical inference to produce robust frame-wise pitch estimates, particularly effective for monophonic signals with soft onsets and vibrato, such as pitchpipe tones.

For ENF frequency estimation, we used the spectral peak of the FFT within a restricted band between 40 and 60 Hz. To maximize frequency resolution, the FFT size was set equal to the full length of the audio segment (typically 30 seconds). This approach provides a frequency bin spacing of approximately 0.033 Hz at a sampling rate of 48 kHz, allowing us to detect subtle deviations from the nominal 50 Hz reference. Prior to the FFT, a Hann window was applied to reduce spectral leakage. The resulting peak frequency was then compared with the mean spectral magnitude within the same 40–60 Hz band to compute a *presence index*. Low index values indicate that the ENF component is absent or buried in noise, while high values suggest a strong and clearly detectable ENF signal.

To evaluate the consistency between pitchpipe and ENF-based estimates of playback speed, we analyzed audio segments containing both signals (illustrated in Figure 4). For each segment, we computed the fundamental frequency of the pitchpipe using the pYIN algorithm and normalized it w.r.t 440 Hz. Similarly, the ENF frequency was estimated via spectral peak analysis and normalized it w.r.t. 50 Hz. Only segments with an ENF presence index greater than 10 were included, ensuring reliable frequency estimation from the noise signal. This threshold also serves to discard recordings likely made with battery-powered equipment, which typically lack a detectable ENF component due to the absence of connection to the electrical grid. Filtering in this way helps avoid spurious correlations and focuses the analysis on segments where both pitchpipe and ENF cues are genuinely present and measurable. As shown in Fig. 5, a strong positive correlation is observed between both measurements, indicating that ENF and pitchpipe provide coherent cues for playback speed estimation. The regression line (in red) closely follows the identity line (dashed), confirming that both sources reflect similar deviations from nominal speed. The Pearson correlation coefficient is 0.85, highlighting a strong linear relationship despite some variability, likely attributable to signal quality or estimation uncertainty.

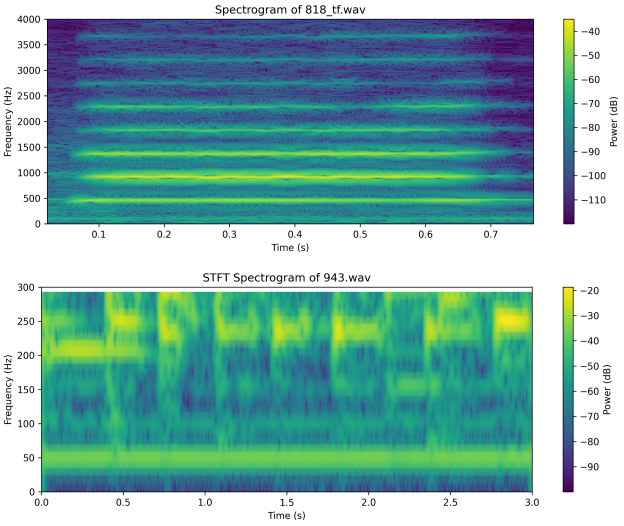


Fig. 4: Spectrograms of audio excerpts from **MagTapeDB**. (a) shows a pitchpipe tone used as a tuning reference, characterized by stable harmonic content. (b) displays a noise-only excerpt in which a 50 Hz ENF component is clearly discernible, enabling playback speed estimation.

Using the fitted regression line, and assuming a nominal ENF frequency of 50 Hz, the average predicted pitchpipe frequency is slightly higher than the standard A440. Specifically, the model suggests a reference frequency close to 447 Hz, which is consistent with historical pitch variability and possible tuning differences in the pitchpipe used during the original recordings. This analysis is an example of how the availability of a curated dataset, combined with tailored signal processing techniques, enables data-driven hypotheses about historical recording practices.

### *B. Denoising*

To demonstrate the usability of **MagTapeDB** with deep learning methods, we implemented a denoising system based on a two-stage U-Net architecture with a Supervised Attention Module (SAM), following the approach proposed in [13],

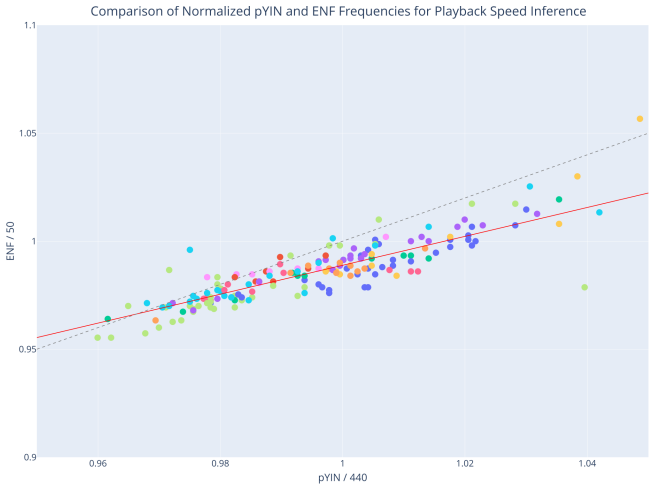


Fig. 5: Correlation between pitchpipe frequency normalized to 440 Hz and ENF frequency normalized to 50 Hz. Each point depicts a matched audio segment containing both signals. Red line = linear regression fit, dashed line = identity line (slope = 1). The strong positive correlation supports the use of ENF as a reliable cue for playback speed estimation. Point color indicates the month and year of the recording.

illustrated in Figure 6. This model has shown strong performance in removing complex background noise from musical recordings while preserving perceptual quality.

Both stages of the model receive as input the complex-valued short-time Fourier transform (STFT) of the noisy signal, where real and imaginary parts are treated as separate input channels, along with a frequency-positional embedding. Given an audio fragment $x$ sampled at 44.1 kHz, the STFT $X$ is computed using a window size of $N = 2048$ and a hop size of $h = 512$ samples.

In the first stage, the network estimates the STFT of the residual noise, denoted as $\hat{Z}$. This estimate is processed by the SAM, which filters and forwards only relevant features to the second stage. The output of the first stage is also used to compute an intermediate estimate of the clean signal via:

$$\hat{Y}_1 = X + \hat{Z}. \tag{1}$$

The second stage then refines this prediction and outputs a final STFT estimate of the clean signal, denoted as $\hat{Y}_2$, using both the first-stage output and the original noisy input. As shown in [13], this two-step design mitigates the risk of musical noise artifacts often introduced by single-stage models.

The model was trained by minimizing the loss function defined in Equation 2, which computes the mean absolute error (MAE) between the clean STFT $Y^k$ and the outputs of both network stages, $\hat{Y}_1^k$ and $\hat{Y}_2^k$, for each time-frequency bin $k$:

$$\mathcal{L} = \frac{1}{K} \sum_k \left( \left| \hat{Y}_1^k - Y^k \right| + \left| \hat{Y}_2^k - Y^k \right| \right), \tag{2}$$

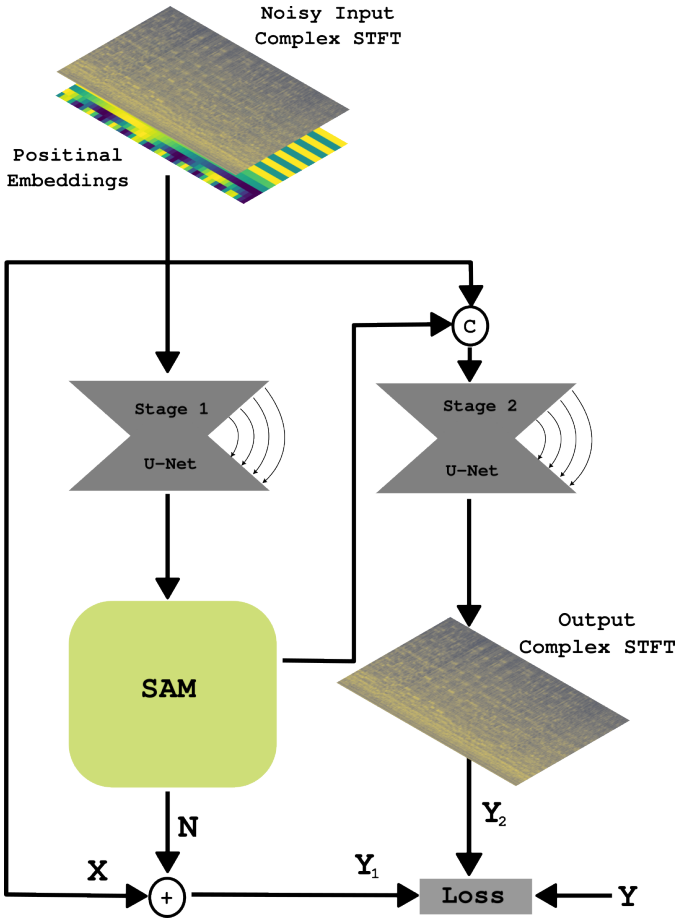


Fig. 6: Two-stage denoising model with Supervised Attention Module (SAM), adapted from [13]. The first stage estimates residual noise and guides the second stage via attention-weighted features. This design aims to reduce musical noise artifacts and improve reconstruction quality.

where $K$ is the total number of STFT bins. This formulation encourages both stages to produce outputs that closely approximate the target clean signal.

Two datasets were used to train the model and evaluate its performance: one containing clean music recordings and another composed of tape noise audio fragments. These were artificially combined to simulate the effect of real tape recordings, by adding tape noise to clean audio. The mixing process is described in Equation 3, where $y$ and $z$ denote clean music and tape noise signals, respectively, and $x$ represents the resulting simulated degraded signal. The parameter $\alpha$ controls the signal-to-noise ratio (SNR), while $\beta$ adjusts a global amplitude scaling factor.

To enhance variability during training, both $\alpha$ and $\beta$ were randomly sampled for each training instance. Specifically, $\alpha$ was drawn from a log-uniform distribution between 6 and 32 dB, and $\beta$ from a log-uniform range between 0 and $-6$ dB.

$$x = \beta(y + \alpha\, z) \tag{3}$$

The tape noise excerpts were sourced from **MagTapeDB**,

and split into 80% for training and 20% for validation. As in [13], clean music signals were taken from the MusicNet dataset [21][4]. MusicNet contains 330 classical music recordings released under open licenses, totaling approximately 34 hours of audio. It is a widely adopted benchmark for supervised learning tasks in music information retrieval. In this work, we adhered to the train/test split proposed in [21], using the designated 1% subset for testing, while partitioning the remaining data into 90% for training and 10% for validation. It is important to note a potential domain shift between the MusicNet corpus —— comprising mainly Western classical music —— and the traditional and folk recordings in **MagTapeDB**. Future work will explore fine-tuning on clean recordings that are musically and acoustically closer to the target domain to better match its timbral and spectral characteristics.

For final evaluation, a test set was constructed using clean musical excerpts from test split of MusicNet and tape noise fragments drawn from an entirely separate dataset [12], to prevent any data leakage. A total of 100 fragment pairs, each 10 seconds long, were mixed at random SNR values between 10 and 20 dB, simulating moderately to highly degraded recording conditions.

The model was trained using the Adam optimizer with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$. The learning rate was initially set to $1 \times 10^{-4}$ and reduced by a factor of 10 every 100,000 steps. Training was performed for a total of 300,000 steps on an NVIDIA RTX 3060 GPU with 12 GB of RAM, taking approximately 48 hours. Early stopping was based on the validation loss, monitored alongside the training loss to avoid overfitting.

The improvement in perceptual quality achieved by the proposed denoising model was quantified using ViSQOL (Virtual Speech Quality Objective Listener) [22] — a full-reference objective metric that estimates perceptual audio quality by comparing time-frequency representations of a clean reference and a degraded signal. In this work, ViSQOL is used to compute the perceptual improvement ($\Delta$MOS) between noisy and denoised signals.

The average $\Delta$MOS obtained across the test set was 1.92, with a standard deviation of 0.96. This result reflects a substantial perceptual enhancement compared to the noisy input signals, demonstrating the model's effectiveness.

## V. Applications and Use Cases

Many of the standard use cases for audio restoration datasets can be addressed using **MagTapeDB**, with the key distinction that this dataset is composed of musicological field recordings, which exhibit their own unique characteristics. These include persistent background noise (e.g., tape hiss), signal saturation, and the presence of environmental sounds typical of in-situ documentation contexts.

In addition to the tasks explored in this work —— such as playback speed estimation and denoising —— the dataset supports a wide range of potential applications. These include instrument detection, performer identification, and inference of recording equipment characteristics (e.g., estimating whether the same recorder was used across sessions). Furthermore, the dataset is well suited for experimenting with saturation mitigation techniques (desaturation), as well as other restoration challenges specific to analog field recordings.

Beyond engineering-oriented applications, **MagTapeDB** also provides researchers in other disciplines —— such as musicology and archival studies —— with access to curated excerpts from a valuable documentary collection. Initiatives of this kind help create the conditions necessary for interdisciplinary collaboration, where shared datasets serve as common ground for joint exploration. Meaningful progress in this domain requires the combined expertise of data scientists, engineers, musicologists, archivists, and other specialists working together to address the technical and cultural challenges of audio preservation.

By making these recordings accessible in a structured and analyzable form, **MagTapeDB** not only enables technological development, but also contributes to the safeguarding of sonic heritage and promotes its use in educational, historical, and artistic contexts.

## VI. Conclusion and Future Work

In this work, we presented **MagTapeDB**, a curated dataset designed to support research in the restoration of historical music recordings stored on magnetic tape. The dataset includes more than 800 annotated audio excerpts comprising musical fragments, isolated tape noise, and pitchpipe tones, along with detailed metadata. We demonstrated how this resource can be used in multiple restoration-related tasks by providing baseline experiments for playback speed correction using ENF analysis and denoising using a deep learning architecture based on a two-stage U-Net.

**MagTapeDB** is freely available for non-commercial research and educational purposes. By making this collection public, we aim to foster reproducible experimentation and contribute to the preservation and analysis of audio heritage.

Future work will focus on expanding the dataset with additional digitized reels, extending the metadata with more detailed annotations (e.g., vocal presence, genre, language), and supporting new tasks such as segmentation, instrument recognition, and quality assessment. We also plan to explore semi-supervised and self-supervised learning approaches that can better leverage the archival nature of the data, especially in low-resource scenarios.


## Acknowledgements

The authors would like to thank the *Centro de Documentación Musical Lauro Ayestarán* (CDM) for providing access to the archival material used in this work. This research was supported by the *Comisión Sectorial de Investigación Científica* (CSIC) of the Universidad de la República, Uruguay. Luiz Biscainho acknowledges the support of the National Council for Scientific and Technological Development (CNPq), Brazil.


[4]Available from Zenodo: https://doi.org/10.5281/zenodo.5120004

## REFERENCES


[1] S. J. Godsill and P. J. Rayner, *Digital Audio Restoration*. Cambridge, United Kingdom: Springer London, 1998.

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

[3] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal processing*, vol. 56, no. 5, pp. 1830–1839, 2008.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[5] A. L. Aditya Arie Nugraha and E. Vincent, "Multichannel audio source separation with deep neural networks," in *IEEE Transactions on audio, speech, and language processing, Vol. 24, No. 9*, Sep. 2016, pp. 1652–1664.

[6] C. V. Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA speech synthesis workshop*, 2016, pp. 159–165.

[7] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496.

[8] P. U. Diehl, Y. Singer, H. Zilly, U. Schönfeld, P. Meyer-Rachner, M. Berry, H. Sprekeler, E. Sprengel, A. Pudszuhn, and V. M. Hofmann, "Restoring speech intelligibility for hearing aid users with deep learning," *Scientific Reports*, vol. 13, no. 1, p. 2719, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-29871-8

[9] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The musdb18 corpus for music separation," 2017. [Online]. Available: http://dx.doi.org/10.5281/zenodo.1117372

[10] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

[11] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *Ismir*, vol. 14, 2014, pp. 155–160.

[12] I. Irigaray, M. Rocamora, and L. W. P. Biscainho, "Noise reduction in analog tape audio recordings with deep learning models," in *AES Int. Conf. on Audio Archiving, Preservation and Restoration*, Culpeper, United States, Jun. 2023, pp. 1–6.

[13] E. Moliner and V. Välimäki, "A two-stage u-net for high-fidelity denoising of historical recordings," in *Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 841–845.

[14] F. Bressan, V. Burini, E. Micheloni, A. Rodà, R. L. Hess, and S. Canazza, "Reading Tapes Backwards: A Legitimate Approach to Saving Time and Money in Digitization Projects?" *Applied Sciences*, vol. 11, no. 15, p. 7092, Jul. 2021, publisher: MDPI AG. [Online]. Available: https://www.mdpi.com/2076-3417/11/15/7092

[15] *ISO/R 16:1955– Standard tuning frequency (Standard musical pitch)*, ISO – International Organization for Standardization Std., 1955.

[16] C. Marvin, "A history of performing pitch: The story of" a"," *Notes*, vol. 60, no. 3, pp. 36–38, 2004.

[17] E. Ngharamike, L.-M. Ang, K. P. Seng, and M. Wang, "Enf based digital multimedia forensics: Survey, application, challenges and future work," *IEEE Access*, vol. 11, pp. 101 241–101 272, 2023.

[18] P. A. Esquef, J. A. Apolinário, and L. W. Biscainho, "Improved edit detection in speech via enf patterns," in *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2015, pp. 1–6.

[19] C. Grigoras, "Applications of enf criterion in forensic audio, video, computer and telecommunication analysis," *Forensic science international*, vol. 167, no. 2-3, pp. 136–145, 2007.

[20] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 659–663.

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

[22] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *Proceedings of the Int. Conf. on Quality of Multimedia Experience (QoMEX)*, Athlone, Ireland, May 2020.