Universidad de la República
Facultad de Ingeniería

Restauración de grabaciones musicológicas mediante técnicas de denoising: Sustracción espectral y Aprendizaje profundo

Memoria de proyecto presentada a la Facultad de Ingeniería de la Universidad de la República por

Analía Arimón, Guillermo Mazzeo, Rodrigo Torrado

en cumplimiento parcial de los requerimientos para la finalización de la carrera de Ingeniería Eléctrica e Ingeniería en Sistemas de Comunicación.

Tutor

Ignacio Irigaray ..... Universidad de la República
Diego Silvera ..... Universidad de la República

Tribunal

Lara Raad ..... Universidad de la República
Pablo Cancela ..... Universidad de la República
Emilio Martínez ..... Universidad de la República

Montevideo
lunes 22 de diciembre, 2025

Restauración de grabaciones musicológicas mediante técnicas de denoising: Sustracción espectral y Aprendizaje profundo, Analía Arimón, Guillermo Mazzeo, Rodrigo Torrado.

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.2).
Contiene un total de 135 páginas.
Compilada el lunes 5 enero, 2026.
http://iie.fing.edu.uy/

El Folklore que por definición es la ciencia que estudia los saberes populares, eso que lleva el hombre, no recibido por vía institucional, sino que por la vía de la tradición, nos hace conocernos, justamente, a nosotros mismos, y ustedes saben muy bien que conocerse a sí mismo es comenzar a mejorarse...

Lauro Ayestarán

Esta página ha sido intencionalmente dejada en blanco.

Resumen

Este trabajo aborda la restauración de grabaciones musicales de cinta analógica mediante técnicas de reducción de ruido (*denoising*), combinando enfoques clásicos de procesamiento de señales con estrategias modernas basadas en aprendizaje profundo. El proyecto surge de la necesidad de preservar el acervo sonoro del *Centro Nacional de Documentación Musical Lauro Ayestarán*, que reúne grabaciones de campo y de estudio afectadas por la degradación propia de los medios magnéticos y las limitaciones tecnológicas de su época.

El objetivo principal consistió en desarrollar, implementar y comparar dos enfoques representativos para la reducción de ruido en grabaciones musicales: (1) un sistema automatizado de *sustracción espectral*, que incluye mejoras orientadas a la restauración musical —como modelado armónico/percusivo, análisis sinusoidal, un esquema iterativo de atenuación progresiva y un detector automático de inactividad de señal—, y (2) un modelo de *aprendizaje profundo* basado en arquitecturas *U-Net* de dos etapas, entrenadas con diferentes bases de datos de ruido (MagTapeDB, con ruido de cinta magnética, y grabaciones de gramófono).

Los entrenamientos se realizaron en el *ClusterUY*, considerando limitaciones de hardware y explorando la influencia del dominio del ruido sobre la capacidad de generalización de los modelos. La evaluación experimental combinó métricas perceptuales objetivas (PEAQ y PAQM), análisis por tipo de contenido musical, variación según la relación señal–ruido (10 dB y 16 dB), tiempos de procesamiento, y escucha crítica cualitativa.

Los resultados demuestran que las técnicas clásicas de procesamiento de señales continúan ofreciendo un rendimiento altamente competitivo. En particular, la sustracción espectral —tanto en su versión estándar como alternativa— logra un equilibrio sólido entre calidad perceptual, estabilidad y eficiencia computacional, manteniendo un desempeño consistente en diversos escenarios, aunque la técnica presente artefactos conocidos como el ruido musical.

Por otro lado, los modelos de aprendizaje profundo muestran un comportamiento más variable: alcanzan resultados competitivos cuando el tipo de ruido y el contenido de las señales coincide con el utilizado en el entrenamiento, pero experimentan una degradación significativa al enfrentarse a dominios no representados. Además, tienden a eliminar transitorios y componentes de alta frecuencia, introduciendo una cierta artificialidad perceptual. Esto evidencia tanto la dependencia de los modelos respecto a los datos de entrenamiento como la limitada explicabilidad de sus decisiones.

Desde el punto de vista práctico, las técnicas basadas en redes neuronales re-

quieren recursos computacionales elevados, tiempos de entrenamiento prolongados y conocimientos especializados para su ajuste y validación, lo cual contrasta con la simplicidad y robustez de los métodos clásicos.

En conjunto, los resultados permiten concluir que las técnicas clásicas siguen siendo una herramienta eficaz y accesible para la restauración de grabaciones patrimoniales, mientras que los enfoques basados en aprendizaje profundo, aunque prometedores, requieren adaptaciones específicas para alcanzar una calidad perceptual comparable en contextos reales y diversos.

Tabla de contenidos

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 1

Introducción

El *denoising* busca atenuar el ruido presente en una señal sin degradar la información sonora relevante. Su importancia radica en que incluso niveles moderados de ruido pueden afectar la inteligibilidad, la percepción musical y el rendimiento de sistemas automáticos basados en audio.

En el caso del habla, la presencia de ruido puede dificultar la comprensión del mensaje, especialmente en ambientes con bajo nivel de señal o con interferencias acústicas significativas. Esto no solo afecta la experiencia del oyente, sino que también puede limitar la eficacia de sistemas automáticos de reconocimiento de voz, asistentes virtuales, o subtitulado automático. La inteligibilidad del habla, es decir, la capacidad de entender las palabras pronunciadas, depende en gran medida de la relación señal-ruido (Signal-to-Noise Ratio, SNR), así como de ciertas características del habla, como los formantes y las consonantes fricativas.

El ruido en grabaciones musicales puede opacar detalles importantes del sonido, como los matices de los instrumentos o la claridad de las voces. Este problema adquiere particular relevancia en el contexto de la preservación del patrimonio musical, donde grabaciones históricas presentan degradaciones características de los medios analógicos que requieren intervención técnica para su restauración.

Este proyecto tiene como objetivo desarrollar, implementar y comparar dos enfoques fundamentales de *denoising* para la restauración de grabaciones musicales: la sustracción espectral con mejoras propuestas y modelos basados en aprendizaje profundo.

La sustracción espectral es una técnica clásica del procesamiento de audio, cuyo principio consiste en estimar el espectro del ruido y sustraerlo del espectro de la señal contaminada. Dado que las implementaciones abiertas disponibles suelen ser limitadas o poco accesibles, parte del proyecto se centró en desarrollar una versión automatizada del algoritmo clásico. Además, se implementó una variante que incorpora modelado espectral mediante técnicas de separación armónica/percusiva y análisis sinusoidal, un esquema iterativo para reducción progresiva del ruido, y algoritmos específicos para mitigación del ruido musical. Un componente clave del sistema desarrollado es el detector automático de inactividad de señal, basado en múltiples métricas (energía en tiempo corto, tasa de cruces por cero y magnitud espectral en altas frecuencias), que permite estimar el perfil de ruido sin intervención

manual.

En paralelo, se entrenaron modelos de *denoising* basados en aprendizaje profundo, empleando una arquitectura *U-Net* de dos etapas para aprender patrones de ruido y restaurar el espectrograma limpio. Los modelos se entrenaron utilizando dos bases de datos de ruidos —MagTapeDB (ruido de cinta magnética) y grabaciones de gramófono—, así como combinaciones de ambas, con el fin de analizar cómo varía su desempeño según el dominio de ruido considerado y evaluar su capacidad de generalización.

La evaluación experimental incorpora múltiples dimensiones: métricas perceptuales objetivas (PEAQ y PAQM), análisis por tipo de contenido musical (música popular, muchas fuentes, pocas fuentes, vocal), variación según relación señal-ruido, tiempos de procesamiento, y escucha crítica cualitativa. Esta evaluación integral permite no solo cuantificar el desempeño técnico de cada enfoque, sino también comprender sus ventajas relativas, limitaciones prácticas y artefactos característicos.

1.1. Motivación

El estudio y desarrollo de técnicas de reducción de ruido en audio responde tanto a necesidades prácticas como científicas. En el ámbito de la restauración de grabaciones musicales históricas, estas técnicas adquieren especial relevancia, ya que permiten recuperar información sonora valiosa afectada por las limitaciones inherentes de los medios analógicos. Sin embargo, muchas de las herramientas disponibles en el entorno profesional son de carácter propietario, presentan un funcionamiento opaco y requieren recursos computacionales elevados o licencias de alto costo, lo que dificulta su adopción en contextos académicos y patrimoniales.

En este escenario, resulta necesario contar con enfoques abiertos, comprensibles y reproducibles que permitan estudiar los mecanismos de reducción de ruido y adaptarlos a distintos tipos de degradación. Este trabajo se propone contribuir en esa dirección, explorando dos paradigmas complementarios: las técnicas clásicas basadas en procesamiento de señales y los modelos modernos de aprendizaje profundo.

La sustracción espectral fue seleccionada como punto de partida por su solidez teórica, su bajo costo computacional y su capacidad de ofrecer control explícito sobre los parámetros de atenuación. Además, su comportamiento y artefactos son bien comprendidos en la literatura, lo que facilita proponer variantes mejoradas y analizar sus efectos. A partir de esta base, se implementó una versión extendida que incorpora modelado armónico/percusivo, análisis sinusoidal y esquemas iterativos de reducción progresiva.

En paralelo, el rápido avance de las redes neuronales profundas abre nuevas posibilidades para la restauración de audio, especialmente en contextos donde el ruido presenta estructuras complejas o no estacionarias. Los modelos tipo *U-Net* han demostrado un desempeño notable en tareas de separación y limpieza de audio.

En definitiva, este trabajo busca aportar una base experimental sólida que permita comprender las ventajas y limitaciones de cada enfoque, y sirva de referencia

para futuros desarrollos en restauración de audio histórico.

1.2. Grabaciones en cinta magnética y su degradación

A mediados del siglo XX, el ruido en las grabaciones analógicas representaba un desafío significativo debido a las limitaciones tecnológicas de la época. Los sistemas de grabación y reproducción utilizaban medios físicos como cintas magnéticas y discos de vinilo, goma, laca o acetato, los cuales eran susceptibles a diversas fuentes de interferencia. El ruido de fondo, a menudo causado por imperfecciones en el medio de grabación, fluctuaciones en la corriente eléctrica, o el desgaste del equipo, se manifestaba como silbidos, zumbidos o distorsiones no deseadas.

Aunque muchas grabaciones profesionales alcanzaban una calidad sonora notable, el ruido seguía presente como una característica inherente del formato, especialmente al realizar copias sucesivas [1]. A diferencia del audio digital, donde las copias pueden ser idénticas al original, en el dominio analógico cada duplicación generaba una pérdida acumulativa de calidad.

Si bien se desarrollaron técnicas para mitigar estos problemas —como el uso de filtros, cintas de mayor calidad y sistemas de reducción de ruido tipo Dolby—, estas soluciones no lograban eliminar completamente las degradaciones, y en algunos casos introducían artefactos propios [2].

Se profundiza sobre estas degradaciones en el Anexo B.2.

1.3. Centro Nacional de Documentación Musical Lauro Ayestarán

Lauro Ayestarán (1913-1966) fue un destacado musicólogo, investigador y docente uruguayo, considerado el pionero de la musicología en el país. Su trabajo fue fundamental para la recopilación, estudio y preservación del patrimonio musical uruguayo, abarcando tanto la música académica como las expresiones musicales populares y folklóricas.

Uno de sus aportes más significativos fue la realización de extensas grabaciones de campo a lo largo de Uruguay, donde documentó diversas manifestaciones musicales tradicionales. Estas grabaciones constituyen un acervo invaluable para la investigación musicológica y la preservación de la cultura sonora del país. Su labor ha sido reconocida internacionalmente, y su legado sigue vigente a través del *Centro de Documentación Musical Lauro Ayestarán* (CDM) [3], que se dedica a la conservación y estudio de sus archivos. En la Figura 1.1 se observan dos momentos de su trabajo de documentación sonora, tanto en entornos de grabación controlados como en el registro de campo.

El Centro Nacional de Documentación Musical Lauro Ayestarán fue creado por resolución del Ministerio de Educación y Cultura de fecha 26 de marzo de 2009, sobre la base de los materiales del archivo del gran musicólogo adquiridos en 2002 por el Estado uruguayo. El proyecto del CDM se basa en el espíritu de Lauro

Figura 1.1: Lauro Ayestarán en distintas instancias de su labor de documentación musical: en estudio y en trabajo de campo, registrando interpretaciones de músicos populares uruguayos mediante grabadores de cinta.

Ayestarán, pionero de una musicología uruguaya, abarcativa de todos los ámbitos de actividad cultural que presentan aspectos musicales, con una visión abierta a otras expresiones culturales, a otros ámbitos antropológicos, a otras manifestaciones artísticas [3].

El trabajo de campo de Lauro Ayestarán se inicia en 1943 y, a partir de 1946, incorpora el registro sistemático en discos de 25 cm a 78 rpm. En 1952 adquiere su primer grabador de cinta magnética, formato que pasa a utilizarse de forma habitual en las campañas posteriores. Algunos años más tarde, hacia 1955, Ayestarán realiza además un respaldo en cinta del material registrado originalmente en disco, de modo que el acervo conservado en el CDM combina registros efectuados directamente en cinta con transferencias posteriores desde soportes de 78 rpm.

En la Fig. 1.2 se observa una cinta magnética de carrete abierto empleados por Ayestarán en estas campañas.

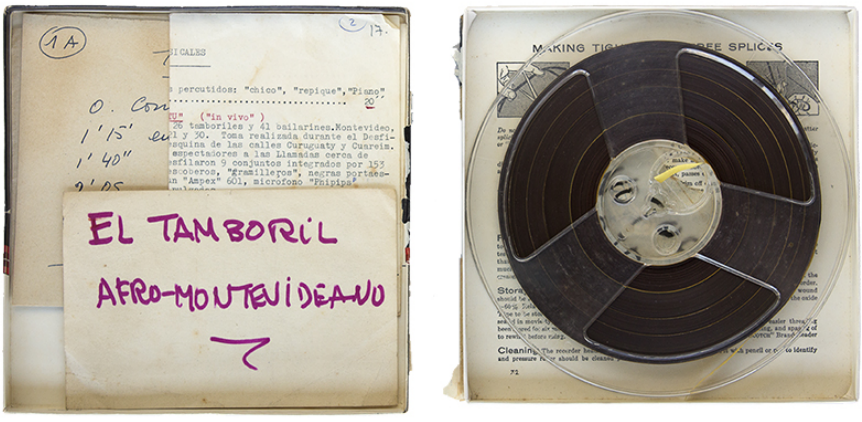

Figura 1.2: Cinta magnética de carrete abierto utilizada por Lauro Ayestarán para el registro sonoro. Imágen del CDM [3].

En este contexto, el presente trabajo se centra explícitamente en el estudio y tratamiento del ruido asociado a las grabaciones en cinta magnética. Esta elección responde, por un lado, a una continuidad natural con el trabajo previo de Irigaray et al. [4], también focalizado en ruido de cinta, y por otro, al objetivo de desarrollar una herramienta con mayor grado de generalidad, que pueda aplicarse no sólo al caso particular de las transferencias desde discos de 78 rpm, sino también a otros archivos sonoros registrados o preservados en cinta magnética. De este modo, la motivación histórica vinculada al archivo de Ayestarán convive con un enfoque metodológico que prioriza la utilidad del método en escenarios más amplios de restauración de audio.

1.4. Antecedentes

El proyecto de *denoising* que se propone tiene como antecedentes varios trabajos y estudios previos que abordan la eliminación de ruido en grabaciones de audio, utilizando tanto técnicas tradicionales como enfoques basados en aprendizaje profundo.

Uno de los antecedentes es el artículo titulado *Aproximación interdisciplinaria al trabajo con documentos sonoros. Estudio de caso: las grabaciones de campo de Lauro Ayestarán* [5], presentado por Ignacio Irigaray y Federico Sallés. Este trabajo aborda la necesidad de re-digitalizar estas grabaciones utilizando tecnologías modernas y procedimientos actualizados de limpieza y digitalización. El proceso incluye la evaluación del estado de los materiales originales, el desarrollo de nuevos algoritmos de procesamiento digital y la implementación de técnicas como la sustracción espectral y eliminación de clicks.

Recientemente, se ha visto un auge en el uso de técnicas de aprendizaje profundo para la reducción de ruido en grabaciones de audio. Un ejemplo es el artículo *A Two-Stage U-Net for High-Fidelity Denoising* [6], donde se trabajó sobre discos de 78 rpm. Otro artículo relevante es *Bandwidth Extension of Historical Music using Generative Adversarial Networks* [7].

Además, se cuenta con un artículo de Ignacio Irigaray, Martin Rocamora y Luiz W. P. Biscainho del 2023, titulado *Noise reduction in analog tape audio recordings with deep learning models* [4], que aborda el problema de la reducción de ruido en grabaciones de cinta utilizando un enfoque de aprendizaje profundo.

Finalmente, existen herramientas comerciales especializadas, como iZotope RX [8], que ofrecen soluciones para la restauración de audio.

1.5. Estructura del documento

El presente trabajo se organiza de la siguiente manera:

- **Capítulo 2**: introduce la técnica de sustracción espectral, desde sus fundamentos teóricos hasta la implementación desarrollada. Se describe el algoritmo clásico, el detector automático de inactividad propuesto y la variante

mejorada *SS Denoisify*, que incorpora modelado espectral, procesamiento iterativo y técnicas de reducción de ruido musical.

- **Capítulo 3**: presenta el modelo de aprendizaje profundo basados en arquitecturas *U-Net* de dos etapas. Se revisan los antecedentes del uso de redes neuronales en procesamiento de audio musical y se detalla el modelo adoptado como base para este trabajo.
- **Capítulo 4**: describe la metodología experimental, incluyendo las bases de datos utilizadas (MusicNet, ruido de cinta magnética, ruido de gramófono y música personalizada), las métricas de evaluación empleadas, el proceso de búsqueda de hiperparámetros para los algoritmos de sustracción espectral, los procedimientos de entrenamiento de los modelos de aprendizaje profundo en el *ClusterUY*, así como la forma en que se llevó a cabo la evaluación final de los modelos propuestos.
- **Capítulo 5**: presenta el análisis de resultados, organizado en cuatro dimensiones: los hiperparámetros óptimos encontrados, las curvas de aprendizaje de los modelos neuronales, la evaluación objetiva mediante métricas perceptuales (considerando el desempeño general, la variación según SNR, el análisis por categoría de contenido y los tiempos de procesamiento), y la escucha crítica cualitativa de las señales restauradas.
- **Capítulo 6**: sintetiza las conclusiones principales del trabajo, identificando las ventajas y limitaciones de cada enfoque, y propone líneas futuras de investigación.

Finalmente, se incluyen dos Apéndices. El Apéndice A analiza métricas alternativas evaluadas para la detección de inactividad, mientras que el Apéndice B caracteriza en detalle los tipos de ruido presentes en discos de gramófono y cintas magnéticas.

Capítulo 2

Sustracción espectral

En el presente capítulo se introduce la primera técnica de *denoising* abordada en este trabajo: la sustracción espectral. Se comienza con una reseña histórica que contextualiza el surgimiento de esta metodología y su relevancia en el procesamiento de señales ruidosas. A continuación, se presenta la formulación clásica del algoritmo y su implementación fundamental.

Posteriormente, se desarrolla en detalle uno de los módulos centrales para su correcto funcionamiento: el detector de inactividad, encargado de identificar de manera autónoma los segmentos de la señal donde no existe actividad relevante y que, por tanto, pueden emplearse para estimar el perfil de ruido. Finalmente, se discuten variantes y extensiones del enfoque tradicional, lo que culmina en el desarrollo de una implementación alternativa propuesta en este trabajo, orientada a mejorar la eficacia del algoritmo en la restauración de grabaciones de audio.

2.1. Introducción

La primera propuesta formal de sustracción espectral fue presentada por Steven F. Boll en 1979, en su artículo *"Suppression of Acoustic Noise in Speech Using Spectral Subtraction"* [9]. Este trabajo dio origen a uno de los métodos más influyentes y ampliamente utilizados en la reducción de ruido en señales de voz [10–12]. Su popularidad se debe principalmente a su sencillez conceptual, bajo costo computacional y facilidad de implementación en tiempo real [10,11,13]. Además, diversas variantes del método han sido incorporadas en sistemas comerciales, incluyendo algoritmos de cancelación de ruido en teléfonos móviles [10].

La técnica propuesta por Boll se fundamenta en que, durante las pausas de habla, la señal registrada está compuesta mayoritariamente por el ruido de fondo. Esto permite estimar su espectro y sustraerlo posteriormente del resto de la señal para obtener una versión más limpia de la voz. El enfoque asume que el ruido es aditivo, independiente y localmente estacionario, de modo que la estimación obtenida en segmentos sin voz se mantiene válida en los instantes inmediatamente posteriores.

A pesar de su efectividad y simplicidad, el método presenta un problema bien

conocido: la generación de ruido musical (*musical noise*), considerado uno de los mayores desafíos de la técnica [9, 10, 13, 14]. Este artefacto se manifiesta perceptualmente como tonos breves, fluctuantes y molestos, resultado de la estructura discontinua que produce la sustracción espectral.

Para mitigar estas distorsiones, se han propuesto numerosas variantes del método original. Entre los aportes más influyentes se encuentra la propuesta de Berouti et al. [14], quienes introdujeron un factor de sobreestimación del ruido y un piso espectral. Otras líneas de trabajo relevantes incluyen la sustracción espectral multibanda [15, 16], los métodos basados en filtrado de Wiener [17], las técnicas iterativas [10, 16, 18–20], los enfoques perceptuales [21], y modelos estadísticos avanzados como los estimadores MMSE de Ephraim y Malah [22].

2.2. Formulación matemática de la técnica

En esta sección se presentan los fundamentos y procedimientos matemáticos que describen la técnica de sustracción espectral [9, 10, 13, 14]. En primer lugar, se considera una señal ruidosa y compuesta por L muestras. El ruido se modela como señal estocástica, aditiva y no correlacionada con la señal determinista original. Bajo este supuesto, puede escribirse:

$$y[m] = x[m] + n[m], \tag{2.1}$$

donde $x[m]$ representa la señal libre de ruido y $n[m]$ una realización del ruido aditivo que la contamina. La transformada de Fourier de tiempo corto (STFT) de la señal y se define como:

$$Y[f,k] = \sum_{m=0}^{L_{\text{fft}}-1} y[m + fL_{\text{hop}}]\, w[m]\, e^{-j\,\frac{2\pi}{L_{\text{fft}}}km}, \tag{2.2}$$

donde $w[m]$ es una ventana de Hann[1] de longitud L_{fft}, que corresponde con el tamaño de la FFT[2], y L_{hop} es el desplazamiento (*hop-size*) entre ventanas consecutivas. El índice f corresponde al número de *frame*, con $f \in \mathcal{F} = \{0, \ldots, L_f - 1\}$, siendo L_f la cantidad total de frames. Por su parte, k indica el *bin* frecuencial y puede tomar los valores $k = 0, \ldots, L_{\text{fft}} - 1$. La cantidad total de frames está dada por:

$$L_f = \left\lfloor \frac{L - L_{\text{fft}}}{L_{\text{hop}}} \right\rfloor + 1. \tag{2.3}$$

[1]La ventana de Hann es una función suavizante que atenúa los bordes de cada segmento para disminuir el efecto de las discontinuidades introducidas por el enventanado y la superposición.

[2]La *Fast Fourier Transform* (FFT) es un algoritmo eficiente para calcular la transformada discreta de Fourier (DFT), reduciendo su complejidad computacional de $\mathcal{O}(N^2)$ a $\mathcal{O}(N \log N)$ y permitiendo obtener el contenido espectral de un *frame* de una señal de manera rápida.

A partir de la Ecuación 2.2, y dado que la STFT es un operador lineal, se obtiene inmediatamente:

$$Y[f,k] = X[f,k] + N[f,k], \tag{2.4}$$

donde $X[f,k]$ corresponde a la STFT de $x[m]$ y $N[f,k]$ a la STFT de $n[m]$.

Para caracterizar el comportamiento espectral, se define el *frame* f de la STFT del ruido como:

$$\mathbf{N}_f = \big[\, |N[f,0]|,\ |N[f,1]|,\ \ldots,\ |N[f,L_{\text{fft}}-1]| \,\big]. \tag{2.5}$$

Cada uno de estos *frames* son realizaciones de un vector estocástico con media $\boldsymbol{\mu}$ y varianza $\boldsymbol{\sigma}^2$.

Dado el escenario matemático anterior, la técnica de sustracción espectral tiene como objetivo atenuar la media introducida por el ruido. El caso ideal para estimar la media espectral del ruido —lo que en este trabajo se denominará **perfil del ruido**— consiste en emplear el siguiente estimador insesgado:

$$\bar{N}[k] = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} N_f[k]. \tag{2.6}$$

Sin embargo, es evidente que este cálculo resulta inviable en la práctica, ya que exige disponer de la STFT del ruido de forma independiente a la STFT de la señal ruidosa; en otras palabras, asumir esto equivale a resolver de antemano el propio problema de la restauración.

Por este motivo, la técnica de sustracción espectral propone modelar el ruido como un proceso *estacionario* a lo largo de toda la señal, es decir, asumir que sus propiedades estadísticas —en particular, su media y su varianza espectral— permanecen aproximadamente constantes en el tiempo. Bajo esta hipótesis, las distintas realizaciones del ruido, observadas en los *frames* donde no hay contenido relevante de la señal, pueden emplearse para estimar de manera consistente su perfil espectral.

Para ello, se asume la existencia de un subconjunto de *frames* $\tilde{\mathcal{F}} \subseteq \mathcal{F}$ en los cuales la señal está ausente y solo se encuentra presente el ruido. En dichos *frames* se cumple:

$$Y[f,k] = N[f,k] \qquad \forall\, f \in \tilde{\mathcal{F}}. \tag{2.7}$$

De este modo, a partir de la condición de estacionariedad del ruido, se define el siguiente estimador del perfil espectral:

$$\hat{N}[k] = \frac{1}{|\tilde{\mathcal{F}}|} \sum_{f \in \tilde{\mathcal{F}}} |Y[f,k]| = \frac{1}{|\tilde{\mathcal{F}}|} \sum_{f \in \tilde{\mathcal{F}}} Y_f[k], \tag{2.8}$$

donde $Y_f[k]$ denota la componente k-ésima del *frame* f de la STFT de la señal ruidosa Y.

La **sustracción espectral** para cada *frame* se define como:

$$\hat{\mathbf{X}}_f = \text{máx}\Big\{ \mathbf{Y}_f - \alpha\, \hat{\mathbf{N}},\ \beta\, \mathbf{Y}_f \Big\}, \qquad f \in \mathcal{F}, \tag{2.9}$$

donde los parámetros $\alpha > 1$, $\beta \in (0, 1)$ controlan, respectivamente, la cantidad de energía espectral sustraída y el piso mínimo permitido para cada componente. Estos valores determinan el compromiso entre la reducción de ruido y la preservación de la calidad de la señal procesada.

Por un lado, el *factor de sobreestimación* α permite incrementar la cantidad de ruido sustraído, multiplicando el espectro estimado del ruido por un valor mayor que uno [9]. Esto atenúa o elimina la mayoría de los picos anchos del espectro de ruido. Sin embargo, en algunas frecuencias permanecen ciertos picos angostos, rodeados por frecuencias de menor potencia, formando lo que [14] denomina *valles*. Estos valles generan transiciones abruptas en el espectro que se perciben auditivamente como el denominado *ruido musical*, el cual se analiza en detalle en la Sección 2.5.

Para suavizar estas transiciones, se introduce el parámetro β, que evita que la magnitud espectral se reduzca abruptamente a cero. De este modo, se atenúan las oscilaciones bruscas entre picos y valles, reduciendo el ruido musical. En la Figura 2.1 se ilustran los resultados de la sustracción espectral aplicada a un *frame* de una señal musical con su correspondiente perfil de ruido. Allí se observan los denominados *valles*, generados por la sustracción espectral sin el parámetro β, y cómo la incorporación del mismo permite suavizar dichas discontinuidades.

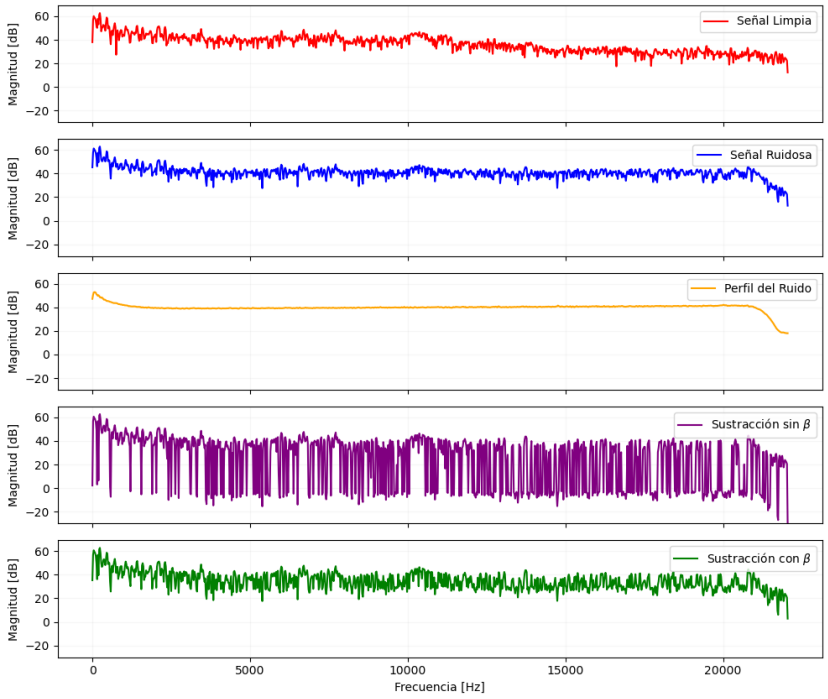

Figura 2.1: Comparación de la sustracción espectral en un *frame* de señal musical. Se muestran: (i) espectro de la señal limpia, (ii) espectro de la señal ruidosa, (iii) perfil de ruido, (iv) resultado de la sustracción espectral con factor de sobreestimación α, donde aparecen los mencionados *valles*, y (v) sustracción espectral reforzada con el parámetro β, el cual suaviza dichas transiciones y reduce las fluctuaciones.

Finalmente, para reconstruir la señal en el dominio temporal se utiliza la magnitud procesada $\hat{X}_f[k]$, obtenida mediante la sustracción espectral, junto con la fase original de la STFT ruidosa $\angle Y[f,k]$. Así, el espectro complejo estimado para cada *frame* se define como:

$$\hat{X}[f,k] = \hat{X}_f[k]\, e^{j\,\angle Y[f,k]}. \tag{2.10}$$

Luego, cada *frame* temporal se obtiene aplicando la transformada inversa de Fourier iFFT:

$$\hat{x}_f[t] = \frac{1}{L_{\text{fft}}} \sum_{k=0}^{L_{\text{fft}}-1} \hat{X}[f,k]\, e^{j\,\frac{2\pi}{L_{\text{fft}}}kt}, \qquad t = 0, 1, \ldots, L_{\text{fft}} - 1. \tag{2.11}$$

La reconstrucción completa (iSTFT) se obtiene mediante el procedimiento de *overlap–add*, sumando las contribuciones de cada *frame* en sus posiciones temporales correspondientes:

$$\hat{x}[m] = \frac{1}{C} \sum_{f=0}^{L_f-1} \Re\{\hat{x}_f[m - fL_{\text{hop}}]\}\ \mathbf{1}_{\{0 \le m - fL_{\text{hop}} < L_{\text{fft}}\}},\ m = 0, 1, \ldots, \tilde{L} - 1, \tag{2.12}$$

donde $\tilde{L} = (L_f - 1)L_{\text{hop}} + L_{\text{fft}}$, y el factor de normalización C se define como:

$$C = \frac{1}{L_{\text{hop}}} \sum_{m=0}^{L_{\text{fft}}-1} w[m]. \tag{2.13}$$

Aquí, $\mathbf{1}_{\{.\}}$ denota la función indicatriz, que toma el valor 1 cuando la condición especificada se cumple y 0 en caso contrario.

Cabe destacar que, en general, no necesariamente se cumple que $L = \tilde{L}$. Por lo tanto, la señal original $x[m]$ y la señal restaurada $\hat{x}[m]$ pueden diferir en su cantidad total de muestras.

2.3. Algoritmo *SS Clásico*

Como punto de partida, se implementó un algoritmo básico de sustracción espectral: *SS Clásico*. La implementación se basó principalmente en el enfoque presentado por Vaseghi en [13], donde se detallan diversos métodos para la reducción de ruido, incluyendo variantes de esta técnica, así como sus fundamentos estadísticos y perceptuales. El esquema general del algoritmo desarrollado se presenta en la Figura 2.2.

Inicialmente, la señal de audio $y[m]$ es transformada al dominio tiempo frecuencia mediante la STFT, utilizando una ventana de tipo Hann. El resultado es la matriz compleja $Y[f,k]$, donde cada columna representa el espectro de un *frame* temporal.

Posteriormente, se aplica un algoritmo de detección de no actividad, cuyo objetivo es identificar segmentos (*frames*) donde no hay contenido musical ni vocal,

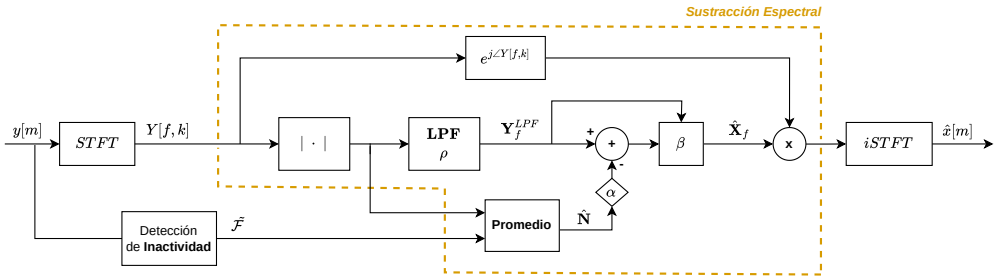

Figura 2.2: Diagrama de bloques del algoritmo básico de sustracción espectral implementado, *SS Clásico*, que incluye el análisis y síntesis STFT, detección de segmentos de inactividad, filtrado pasabajos y la sustracción con parámetros α y β.

sino únicamente ruido ($\tilde{\mathcal{F}}$). Vale la pena recordar que la detección *automática* de estas regiones es uno de los objetivos principales del presente trabajo, ya que permite adaptar el algoritmo de forma dinámica a diferentes entornos de ruido sin una intervención manual. Esto será desarrollado con mayor profundidad en la Sección 2.4.

Los espectros de los segmentos identificados como ruido se promedian para obtener un perfil espectral estimado $\hat{\mathbf{N}}$, que representa el módulo promedio del ruido por cada *bin* de frecuencia sobre el conjunto $\tilde{\mathcal{F}}$ de *frames*, definido anteriormente.

Como se describió en la formulación matemática de la técnica, la presencia de ruido estocástico aditivo sobre una señal modifica su distribución estadística. En particular, la media y la desviación estándar de la señal original se ven afectadas. El estimador propuesto por S. F. Boll [9] plantea corregir la alteración en la media introducida por el ruido a través de la sustracción, pero no compensa la dispersión (varianza) de la señal.

Por esta razón, en [13] se propone aplicar un filtro pasabajos de primer orden (*low-pass filter*, LPF) en la dimensión temporal, con el objetivo de atenuar la varianza no deseada introducida por el ruido. Para ello, se toma el módulo de la STFT $Y[f,k]$ y luego se procesa cada uno de sus *frames* mediante la siguiente función recursiva:

$$\mathbf{Y}_f^{\mathrm{LPF}} = \begin{cases} \mathbf{Y}_0, & f = 0, \\ \rho\, \mathbf{Y}_{f-1}^{\mathrm{LPF}} + (1-\rho)\, \mathbf{Y}_f, & f > 0, \end{cases}$$

donde $\rho \in (0,1)$ es un parámetro que controla la suavidad del filtrado: valores cercanos a 1 producen un filtrado más agresivo y, por lo tanto, una mayor atenuación de las variaciones rápidas.

Si bien una señal de audio puede presentar variaciones abruptas en el tiempo —por ejemplo, debido a transiciones rápidas entre formantes o a cambios súbitos en su estructura espectral— se asume que estas variaciones son menos pronunciadas que las generadas por el ruido analógico.

Bajo esta hipótesis, el uso de un LPF en el tiempo implica un compromiso: se acepta cierto riesgo de suavizar componentes legítimas de la señal a cambio de

suprimir las fluctuaciones más rápidas asociadas al ruido. Según [13], este suavizado espectral contribuye a reducir la varianza del ruido y, por ende, a mitigar la aparición de ruido musical.

Finalmente, se aplica la sustracción espectral sobre los *frames* resultantes del filtrado pasabajos, siguiendo el mismo procedimiento definido en la Ecuación 2.9. La señal en el dominio temporal se reconstruye luego mediante la iSTFT, empleando como espectro complejo el producto entre la magnitud estimada $\hat{X}[f,k]$ y la fase de la STFT ruidosa, esto es, $\hat{X}[f,k]\, e^{j\angle Y[f,k]}$.

2.4. Detector de inactividad de la señal

En la técnica de sustracción espectral es fundamental disponer de un perfil de ruido representativo, obtenido a partir de segmentos de la señal donde no existe actividad útil. Tradicionalmente, esta identificación de regiones inactivas se realizó de forma manual, lo que limita la autonomía y escalabilidad del proceso.

En este trabajo se propone un enfoque completamente automático para la detección de inactividad, cuyo objetivo es localizar de manera robusta los tramos libres de contenido relevante y, a partir de ellos, estimar el perfil de ruido sin una intervención manual. Este módulo constituye una etapa clave dentro del sistema de *denoising*, ya que la calidad de la estimación del ruido condiciona directamente el desempeño de la sustracción espectral.

A continuación se detalla la implementación completa del algoritmo propuesto para la detección automática de inactividad. Dado que las decisiones de diseño adoptadas en este módulo se fundamentan en el comportamiento de las métricas evaluadas, se recomienda revisar previamente el análisis presentado en el Apéndice A, donde se discuten en profundidad las propiedades, ventajas y limitaciones de cada métrica considerada.

Implementación del algoritmo

En primer lugar, el detector de inactividad analiza la señal en el dominio temporal mediante ventanas solapadas (*frames* temporales), utilizando el mismo tamaño de ventana y el mismo desplazamiento que los utilizdos en la STFT. Es importante recordar que el funcionamiento del algoritmo implementado se basa en la hipótesis de que el ruido presente en la grabación es aproximadamente estacionario a lo largo de toda su duración. Bajo este supuesto, se asume que los últimos *frames* de la señal contienen únicamente ruido, lo cual permite obtener una primera estimación de su perfil. Esta elección se justifica en que, típicamente, las piezas musicales y las grabaciones musicológicas no finalizan de forma abrupta, sino que incluyen una breve sección final sin contenido musical relevante, que puede aprovecharse como referencia inicial para caracterizar el ruido presente en toda la señal.

Posteriormente, se aplican umbrales específicos sobre cada una de las métricas extraídas, con el objetivo de detectar de forma robusta los segmentos de inactividad. Para evitar clasificaciones erróneas causadas por transiciones graduales entre regiones activas e inactivas —como los ataques o decaimientos al inicio o

final de un sonido— se incorporan márgenes adicionales al comienzo y al final de cada segmento detectado como silencio. Estos márgenes permiten excluir los *frames* limítrofes que podrían estar contaminados por contenido mixto de sonido y ruido. Por defecto, el margen inicial es mayor que el final, ya que el crecimiento de amplitud al inicio de un sonido suele ser más abrupto que su decaimiento.

Además, el algoritmo impone restricciones de duración mínima tanto para los segmentos de silencio como para los de actividad. Un segmento es considerado silencio únicamente si su longitud excede un umbral mínimo, lo cual previene detecciones falsas provocadas por fluctuaciones breves en las métricas. De forma similar, si se detecta un sonido entre dos silencios cuya duración no alcanza el umbral mínimo para el sonido, se lo considera parte del silencio anterior, evitando así la fragmentación innecesaria de los tramos inactivos.

Inicialmente, para la implementación del algoritmo se consideraron dos métricas: la **energía en tiempo corto** (STE) y la **taza de cruces por cero en tiempo corto** (ZCR). Ambas se calculan aplicando una ventana deslizante de tamaño fijo sobre la señal ruidosa en el dominio temporal, avanzando con un salto definido. En el caso de la STE, se estima la energía de cada *frame* como la suma de los cuadrados de las muestras contenidas en la ventana, de acuerdo con la siguiente expresión:

$$\mathrm{STE}[n] = \sum_{m=0}^{M-1} x^2[nR + m] \tag{2.14}$$

donde $x[n]$ es la señal, $M = L_{\mathrm{fft}}$ es el tamaño de la ventana y $R = L_{\mathrm{hop}}$ el salto entre ventanas. Por otro lado, la ZCR se calcula contando cuántas veces la señal cambia de signo dentro de cada ventana, lo cual se obtiene a partir del signo de muestras consecutivas. El resultado se normaliza por el tamaño de la ventana, lo que conduce a la siguiente expresión:

$$\mathrm{ZCR}[n] = \frac{1}{M} \sum_{m=1}^{M-1} \left|\mathrm{sgn}(x[nR + m]) - \mathrm{sgn}(x[nR + m - 1])\right| / 2 \tag{2.15}$$

donde $\mathrm{sgn}(\cdot)$ representa la función signo.

No obstante, como se discute en el Apéndice A, es fundamental tener en cuenta los casos en los que ciertos fragmentos de la señal, particularmente aquellos con componentes agudas, puedan presentar una tasa de cruces por cero elevada sin corresponder necesariamente a ruido. Para abordar esta situación, se introdujo una nueva métrica: la **magnitud espectral promedio en altas frecuencias** (MHF). Esta métrica se calcula a partir de la STFT de la señal, utilizando el mismo valor de ventana que en las métricas anteriores. En cada *frame*, se computa el promedio de la magnitud espectral a partir de una frecuencia umbral f_{cut}, con el objetivo de estimar la presencia de contenido espectral en altas frecuencias. La fórmula utilizada para calcular esta métrica es:

$$\mathrm{MHF}[n] = \frac{1}{K - k_c} \sum_{k=k_c}^{K-1} |X[n, k]| \tag{2.16}$$

donde $X[n, k]$ representa el valor complejo de la STFT en el *frame* n y bin k, k_c es el bin de frecuencia correspondiente a f_{cut}, y $K = L_{\text{fft}} \ / \ 2$ es el total de bins de frecuencia.

Para determinar la frecuencia de corte f_{cut} y su correspondiente índice espectral k_c, se tomó como referencia el comportamiento de una señal sinusoidal, así como la tasa de cruces por cero promedio en los últimos *frames* de la señal, de los cuales se asume que contienen únicamente ruido. En particular, una sinusoide de frecuencia f, muestreada a una frecuencia f_s, tiene una tasa de cruces por cero dada por:

$$\text{ZCR}_{\text{sin}} = \frac{2f}{f_s} \tag{2.17}$$

A partir del cálculo de la ZCR para cada ventana de análisis, se estima la media sobre los últimos *frames* de silencio, que se denota como $\overline{\text{ZCR}}_{\text{noise}}$. Este valor permite estimar una frecuencia de referencia cuya tasa de cruces por cero sea equivalente, despejando de la ecuación anterior:

$$f_{\text{ref}} = \frac{\overline{\text{ZCR}}_{\text{noise}} \cdot f_s}{2} \tag{2.18}$$

Con el fin de introducir un margen de tolerancia, se define la frecuencia de corte como una fracción de esta frecuencia de referencia:

$$f_{\text{cut}} = (1 - \alpha_{\text{pct}}) \cdot f_{\text{ref}} \tag{2.19}$$

donde $\alpha_{\text{pct}} \in (0, 1)$ es un parámetro de tolerancia que define cuán estricta será la exclusión de componentes de frecuencia inferior. Finalmente, el índice espectral correspondiente a esta frecuencia de corte se obtiene como:

$$k_c = \left\lfloor \frac{f_{\text{cut}} \cdot M}{f_s} \right\rfloor . \tag{2.20}$$

Los umbrales utilizados para cada una de las métricas fueron definidos en función de su valor promedio estimado sobre los últimos *frames*. Para cada métrica, el umbral se establece como una fracción de su media en esta región de referencia, según las siguientes expresiones:

$$T_{\text{STE}} = (1 + \alpha_{\text{STE}}) \cdot \overline{\text{STE}}_{\text{noise}} \tag{2.21}$$

$$T_{\text{ZCR}} = (1 - \alpha_{\text{ZCR}}) \cdot \overline{\text{ZCR}}_{\text{noise}} \tag{2.22}$$

$$T_{\text{MHF}} = (1 + \alpha_{\text{MHF}}) \cdot \overline{\text{MHF}}_{\text{noise}} \tag{2.23}$$

donde $\alpha_{\text{STE}},\ \alpha_{\text{ZCR}},\ \alpha_{\text{MHF}} \in [0, 1]$ son parámetros de sensibilidad que determinan qué tan estrictos serán los umbrales respecto a la energía, la tasa de cruces por cero y la magnitud espectral en altas frecuencias, respectivamente.

Una vez definidos los umbrales, se procede a recorrer todos los *frames*, clasificando como inactivos aquellos que cumplan simultáneamente las siguientes tres condiciones:

$$\begin{aligned} \text{STE}[n] &< T_{\text{STE}} \\ \text{ZCR}[n] &> T_{\text{ZCR}} \\ \text{MHF}[n] &< T_{\text{MHF}} \end{aligned} \tag{2.24}$$

Las condiciones 2.24 aseguran que un *frame* será considerado como inactivo si presenta baja energía, alta tasa de cruces por cero y baja magnitud espectral en altas frecuencias.

En el Algoritmo 1 se puede apreciar la implementación para la detección de inactividad de la señal. La máscara de segmentos inactivos utilizada en este algoritmo es un arreglo del mismo tamaño que la cantidad total de *frames*, donde cada elemento indica si el *frame* correspondiente está activo o inactivo.

Algoritmo 1 Detección de segmentos de inactividad de la señal.

Entrada: Señal x y su frecuencia de muestreo; parámetros: tamaño de ventana, salto entre ventanas, control de umbrales, márgenes, largos mínimos de segmentos, cantidad de ventanas iniciales.

Salida: Máscara de segmentos inactivos.

1. Inicialización

Calcular la magnitud de la STFT de la señal x.
Calcular: STE, ZCR, $\overline{\text{STE}}_{\text{ruido}}$, $\overline{\text{ZCR}}_{\text{ruido}}$.
Calcular: MHF, $\overline{\text{MHF}}_{\text{ruido}}$.
Calcular: T_{STE}, T_{ZCR}, T_{MHF}.
Inicializar la máscara de segmentos inactivos con las ventanas iniciales.

2. Detección de Inactividad

for *n: resto de ventanas* **do**
  if $STE[n] < T_{\text{STE}}$ ***and*** $ZCR[n] > T_{\text{ZCR}}$ ***and*** $MHF[n] < T_{\text{MHF}}$ **then**
    Marcar ventana como inicio de segmento de inactividad.
  else if *se detectó un inicio y el segmento es suficientemente largo* **then**
    Aplicar márgenes.
    Verificar el tamaño del segmento activo entre los segmentos inactivos.
    Actualizar máscara de segmentos inactivos.

3. Postprocesamiento

Procesar el último segmento de silencio (si corresponde).

Retornar *máscara de segmentos inactivos*

2.5. Propuestas de mejora del algoritmo básico

La sustracción espectral, desde su formulación original propuesta por Boll, se consolidó rápidamente como una técnica simple y eficiente para la reducción de ruido en señales de audio, especialmente de voz. Sin embargo, su aplicación práctica evidenció limitaciones importantes, entre ellas la aparición del denominado *ruido musical* y la incorporación de ciertas distorsiones cuando la sustracción es demasiado agresiva y atenúa componentes relevantes de la señal.

Estas dificultades motivaron el desarrollo de diversas variantes orientadas a mejorar la robustez y la calidad perceptual del método. Entre ellas se destacan las estrategias específicas para mitigar el ruido musical —como la eliminación de componentes espectrales de muy baja magnitud o el uso de sustracción espectral

iterativa— y la incorporación de etapas de modelado espectral, destinadas a preservar la estructura relevante de la señal antes y después del proceso de atenuación.

En esta sección se describen las técnicas empleadas para abordar estas limitaciones y mejorar el desempeño del algoritmo básico de sustracción espectral, desarrollado en la Sección 2.3.

2.5.1. Ruido musical

El fenómeno conocido como ruido musical se refiere a un conjunto de artefactos tonales que aparecen en señales procesadas mediante algoritmos de reducción de ruido basados en la sustracción espectral. En una representación tiempo–frecuencia, estos artefactos se manifiestan como picos breves e irregulares, distribuidos aleatoriamente en ambas dimensiones y con mayor predominancia en las bandas altas de frecuencia.

En la Figura 2.3, que muestra el espectrograma de una señal restaurada con el algoritmo *SS Clásico*, pueden identificarse como pequeños picos de color celeste que sobresalen del fondo azul oscuro, el cual corresponde a los valles espectrales donde la energía es considerablemente menor.

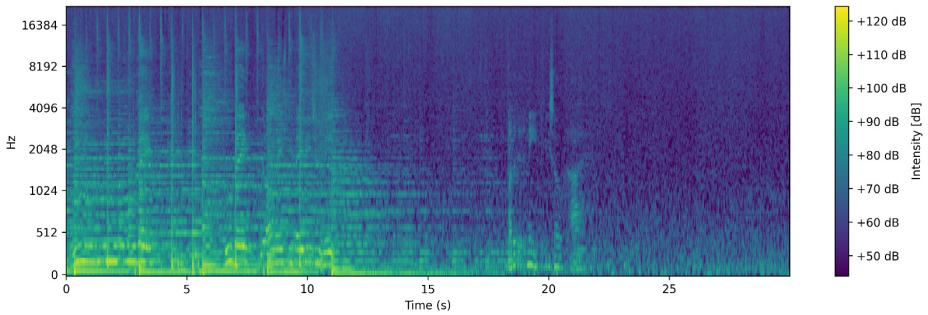

Figura 2.3: Espectrograma de una señal restaurada mediante *SS Clásico*, donde se observan picos espectrales breves e irregulares —característicos del ruido musical— que sobresalen del fondo de baja energía (valles).

Perceptualmente, estos picos no guardan relación con la estructura armónica de la señal original, por lo que se perciben como tonos breves y fluctuantes. Steven F. Boll [9] señala que estos eventos pueden aparecer incluso en regiones donde existe actividad relevante, especialmente cuando la señal no logra enmascararlos, siendo más notorios en regiones de silencio o baja energía.

Diversos trabajos [9, 10, 13, 14] atribuyen la aparición del ruido musical a dos factores principales. En primer lugar, debido al procedimiento descrito en la Ecuación 2.9 que atenúa aquellos coeficientes espectrales cuya energía se encuentra por debajo de la estimación del perfil de ruido, en un factor controlado por la variable β. Este mecanismo genera valles espectrales —huecos abruptos en la distribución del espectro— y produce una representación tiempo–frecuencia irregular y discontinua, tal como se ilustra en la figura anterior.

En segundo lugar, desde una perspectiva estadística, el fenómeno está asociado a la varianza del ruido. La estimación del perfil espectral representa típicamente el

valor promedio del ruido, de modo que los coeficientes que superan aleatoriamente dicha media no son eliminados y permanecen en el residuo. Estas fluctuaciones estocásticas originan dichos picos intermitentes. Dado que este residuo no siempre queda enmascarado por la energía armónica de la señal, sus componentes sobresalientes se vuelven audibles durante la reconstrucción temporal, especialmente en las bandas de frecuencia más altas.

A partir del análisis teórico anterior, se presentan a continuación dos técnicas implementadas con el propósito de mitigar la presencia de ruido musical en la restauración de grabaciones de audio.

Algoritmo de reducción de ruido musical

En este trabajo se implementó una función de detección y supresión de ruido musical que opera en el dominio espectral mediante la STFT. El algoritmo analiza *frame* por *frame* la evolución temporal de los coeficientes espectrales y aplica un proceso de eliminación selectiva: si un componente presenta simultáneamente una duración breve y una magnitud reducida (por debajo de un cierto umbral), se clasifica como ruido musical y su magnitud es anulada. La fase original se conserva y la señal se reconstruye mediante la iSTFT. El procedimiento completo se presenta en el Algoritmo 2.

Algoritmo 2 Algoritmo de atenuación del ruido musical basado en detección espectro-temporal de eventos de baja energía y corta duración.

Entrada: Señal x; parámetros: tamaño FFT N_{FFT}, salto H, umbral en dB T_{dB}, duración máxima permitida $L_{\max}$.

Salida: Señal y con el ruido musical atenuado.

1. Análisis STFT

Calcular la STFT de x: $X \leftarrow \text{STFT}(x, N_{\text{FFT}}, H)$.

Separar magnitud $M = |X|$ y fase $\Phi = \angle X$.

2. Conversión de umbral

Convertir el umbral de dB a escala lineal: $T_{\text{lin}} = 10^{T_{\text{dB}}/20}$.

3. Detección de eventos de baja energía

Construir máscara binaria B: $B[f, n] = 1$ si $M[f, n] < T_{\text{lin}}$, en otro caso 0.

4. Eliminación de eventos cortos

for *cada frecuencia* f **do**

  Detectar inicios y finales de secuencias consecutivas con $B[f, :] = 1$.

  Calcular la longitud $L = \text{fin} - \text{inicio}$.

  if $0 < L \leq L_{\max}$ **then**

    Anular: $M[f, \text{inicio} : \text{inicio} + L] \leftarrow 0$.

5. Reconstrucción

Obtener y mediante $\text{iSTFT}(M \cdot e^{j\Phi}, N_{\text{FFT}}, H)$. **Retornar** y

La Figura 2.4, extraída de [13], ilustra gráficamente este procedimiento: se utiliza una ventana temporal deslizante sobre la magnitud espectral de cada *frame*, comparando su duración y nivel con un umbral predefinido. Los componentes que no cumplen estos criterios son eliminados (marcados con una ×), mientras que aquellos considerados válidos se conservan (marcados con un ✓).

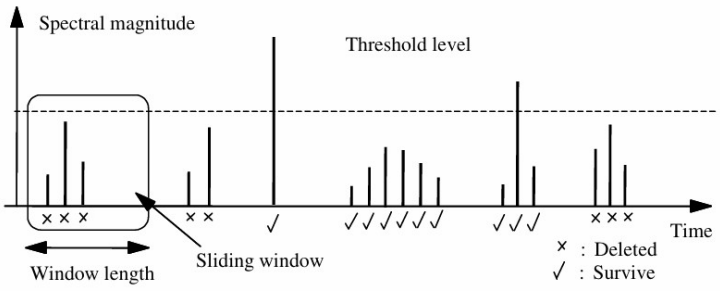

Figura 2.4: Ejemplo del procedimiento de identificación y supresión de ruido musical a partir de características espectro-temporales. La magnitud espectral se recorre con una ventana deslizante, comparando cada evento con un umbral de energía y una duración máxima. Los componentes descartados se marcan con una ×, mientras que los preservados aparecen con un ✓. Imagen extraída de [13].

Sustracción espectral iterativa

Por otro lado, el trabajo *Reinforced Spectral Subtraction Method to Enhance Speech Signal* [10] propone una versión iterativa de la sustracción espectral cuyo propósito es adaptar dinámicamente el perfil de ruido en función del residuo generado en cada etapa. Bajo este enfoque, el ruido musical —aunque no estacionario— se modela como un nuevo ruido aditivo que puede estimarse y atenuarse progresivamente. El método inicia con una sustracción espectral convencional; la señal obtenida, que aún contiene ruido musical, se analiza en segmentos sin presencia musical para estimar el espectro de dicho residuo. Esta estimación se emplea en una nueva sustracción sobre la señal procesada, y el procedimiento se repite de manera iterativa, refinando en cada paso la caracterización del ruido y permitiendo una adaptación continua del algoritmo.

Además, dado que el ruido musical presenta variaciones temporales, Ogata [10] propone realizar estimaciones locales del perfil de ruido en marcos temporales separados para cada iteración, lo que mejora la capacidad de seguimiento de la estructura del ruido residual. El autor reporta resultados considerablemente satisfactorios con esta metodología, mostrando mejoras claras en la calidad de las señales procesadas bajo distintos escenarios de ruido.

La sustracción espectral ha sido comparada ampliamente con el filtro de Wiener [13, 16, 22]. En particular, en [16] se analiza cómo, bajo ciertas condiciones, la versión iterativa del método puede aproximarse progresivamente al comportamiento del filtro de Wiener. A medida que la estimación del ruido se vuelve más precisa y la señal procesada se asemeja a la señal limpia, la función de ganancia utilizada en la sustracción espectral tiende a converger hacia una forma cercana a

la del filtro de Wiener, el cual es óptimo en el sentido del error cuadrático medio (*mean square error*, MSE). Si bien dicha convergencia no se garantiza en todos los casos, este análisis aporta una justificación teórica relevante para el uso de esquemas iterativos.

2.5.2. Modelado espectral

Una posible forma de mejorar la eficacia de la sustracción espectral consiste en preservar, antes de aplicar el proceso de sustracción, aquellas componentes que resultan relevantes para la estructura de la señal. La idea central de la propuesta es que, si se logra separar o modelar adecuadamente la porción útil de la señal—por ejemplo, sus componentes armónicas o transitorias—, entonces la sustracción puede concentrarse casi exclusivamente en atenuar las componentes del ruido. De este modo, se evita alterar el contenido relevante de la señal original y se reduce la probabilidad de introducir distorsiones o artefactos durante la restauración.

Para llevar a cabo esta idea, se empleó el **modelado espectral**, una técnica de procesamiento digital de señales que representa una señal —particularmente de audio— como la combinación de componentes de distinta naturaleza. Formalmente, una señal x puede representarse como la suma de las siguientes tres componentes principales:

$$x = x_s + x_t + x_e, \tag{2.25}$$

donde:

- x_s: corresponde con la componente *sinusoidal*, que representa la parte tonal de la señal. Se modela como la suma de sinusoides con frecuencia, amplitud y fase variables en el tiempo.
- x_t: se define como la componente *transitoria*, que captura eventos abruptos de corta duración, como ataques de instrumentos o consonantes plosivas.
- x_e: es la componente *estocástica*, que representa la energía no armónica o aleatoria, incluyendo consonantes fricativas, ruido ambiental u otras fluctuaciones no estructuradas.

La propuesta se centra en desarrollar un método que permita extraer dichas componentes de la señal original x a partir de la señal ruidosa $x+n$, de manera que la sustracción espectral no las distorsione y se enfoque únicamente en atenuar las componentes estocásticas del ruido n. Para ello, se realizó una revisión bibliográfica con el objetivo de identificar las herramientas más adecuadas para llevar a cabo esta tarea. A continuación, se describen los trabajos más relevantes que se consideraron para el presente trabajo.

En primer lugar, los estudios de McAulay y Quatieri [23], junto con los de Serra [24–27], ofrecieron aportes relevantes que han influido en la evolución del modelado espectral. Por un lado, McAulay y Quatieri propusieron una técnica basada en una representación sinusoidal, donde la señal de voz se descompone en componentes cuya frecuencia, amplitud y fase varían suavemente en el tiempo, permitiendo una reconstrucción precisa incluso en entornos ruidosos. Posteriormente,

Serra extendió este enfoque mediante el modelo *Spectral Modeling Synthesis* (SMS), al introducir una descomposición más general en componentes deterministas (sinusoides) y estocásticas (ruido), lo cual permitió una mayor calidad de síntesis y una mayor flexibilidad en la transformación de señales complejas.

Una implementación destacada de este modelo es `SMS Tools` [28], desarrollada por el *Music Technology Group* de la *Universitat Pompeu Fabra*, bajo la dirección del mismo Xavier Serra. Este conjunto de herramientas de código abierto permite el análisis, transformación y síntesis de señales de audio basándose en el modelo de descomposición determinista-estocástico propuesto por Serra y Smith [25] y, también, en las publicaciones [24, 26, 27, 29].

Por otro lado, en paralelo a estos avances, se desarrollaron técnicas orientadas a la separación de componentes dentro de señales de audio complejas. Una de estas técnicas se denomina *Harmonic/Percussive Source Separation* (HPSS) propuesta por Fitzgerald [30], quien introduce un método simple y eficiente basado en el filtrado por mediana aplicado sobre el espectrograma de la señal. El enfoque se fundamenta en la observación de que las componentes armónicas se manifiestan como estructuras horizontales en el dominio tiempo-frecuencia, mientras que las percusivas aparecen como estructuras verticales. Mediante la aplicación de filtros de mediana en direcciones temporales y frecuenciales, se obtienen representaciones separadas que permiten generar máscaras para aislar cada tipo de componente, como se ilustra en la Figura 2.5. Además, en los *FMP Notebooks* del laboratorio AudioLabs Erlangen [31] se puede encontrar una implementación didáctica y extensible de este enfoque.

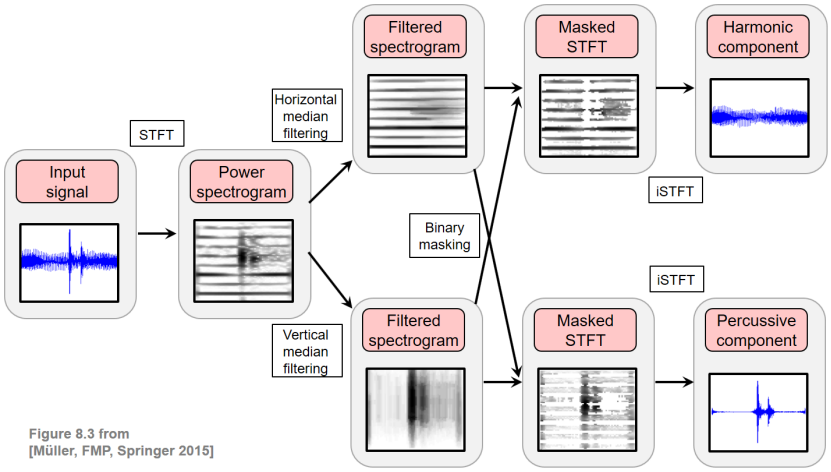

Figura 2.5: Esquema del algoritmo *Harmonic/Percussive Source Separation* (HPSS) propuesto en [31]. A partir del espectrograma de potencia de la señal se aplican filtros de mediana en dirección horizontal y vertical, lo que permite resaltar las estructuras asociadas a componentes armónicas y percusivas, respectivamente. Posteriormente, mediante enmascaramiento binario e iSTFT, se reconstruyen las señales correspondientes a cada tipo de componente.

2.6. Algoritmo *SS Denoisify*

A partir de las técnicas descritas previamente, se diseñó el algoritmo *SS Denoisify*, ilustrado en la Figura 2.6, con el objetivo de mejorar tanto el rendimiento como la eficiencia del método de reducción de ruido basado en sustracción espectral presentado en la Sección 2.3.

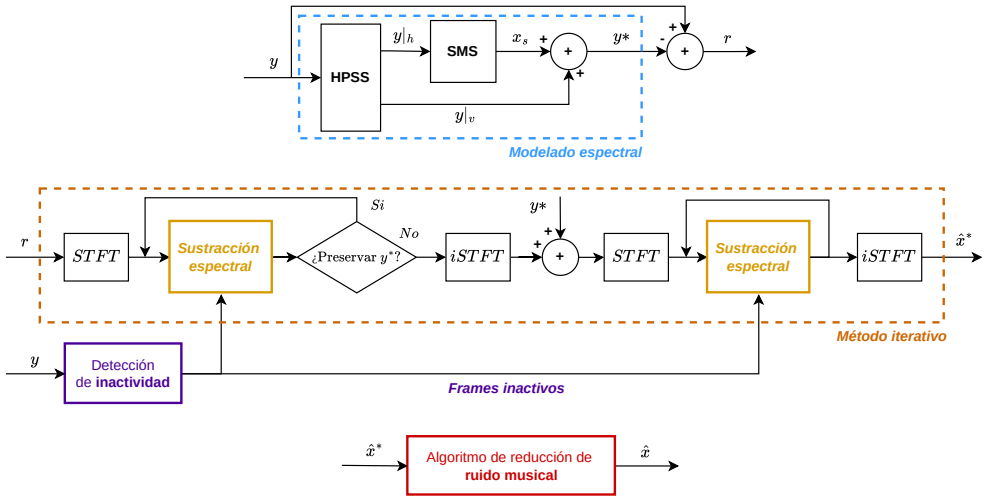

Figura 2.6: Diagrama de bloques del algoritmo *SS Denoisify* propuesto para la reducción de ruido. El proceso combina separación armónica/percusiva (`HPSS`), modelado sinusoidal (`SMS Tools`) y un esquema de sustracción espectral iterativa. Además, incorpora detección de inactividad para estimar el perfil de ruido y una etapa final de supresión de ruido musical.

Inicialmente, se buscó separar las estructuras de la señal que contienen información relevante, como las componentes tonales, los armónicos y los transitorios. Para ello, se empleó la técnica HPSS [31], que descompone la señal ruidosa y en dos: una señal obtenida mediante el filtrado de mediana vertical, $y|_v$, y otra mediante el filtrado de mediana horizontal, $y|_h$, donde se verifica que

$$\text{HPSS}\{y\} = \{y|_v,\ y|_h\}, \quad y = y|_v + y|_h.$$

En este caso, si se asume que la señal ruidosa puede expresarse como

$$y = x + n = x_s + x_t + x_e + n,$$

donde x_s, x_t y x_e son las componentes correspondientes con el modelado espectral, entonces las señales resultantes de los filtrados de mediana pueden escribirse de la siguiente forma:

$$\begin{cases} y|_v = x_s|_v + x_t|_v + x_e|_v + n|_v, \\ y|_h = x_s|_h + x_t|_h + x_e|_h + n|_h. \end{cases} \tag{2.26}$$

Si se considera además que la componente sinusoidal puede obtenerse en su totalidad a partir del filtrado de mediana horizontal [30], $x_s|_h = x_s$, y que la

componente transitoria se puede extraer del filtrado de mediana vertical [30], $x_t|_v = x_t$, se obtiene:

$$\begin{cases} y|_v = x_t + x_e|_v + n|_v, \\ y|_h = x_s + x_e|_h + n|_h. \end{cases} \tag{2.27}$$

Como se observa en la expresión anterior, el ruido puede presentar tanto características armónicas como transitorias, de modo que ambas salidas generadas por el algoritmo HPSS estarán contaminadas por dicho ruido. Por esta razón, sobre la componente $y|_h$ se aplica además un análisis de modelado sinusoidal, con el objetivo de identificar las sinusoides estables a lo largo del tiempo, idealmente asociadas al contenido tonal, melódico o estructurado de la señal. Sea SMS la función encargada del modelado sinusoidal; entonces, se obtiene la siguiente expresión:

$$\text{SMS}\{y|_h\} = \text{SMS}\{x_s + x_e|_h + n|_h\} = x_s. \tag{2.28}$$

Para ilustrar este proceso, en la Figura 2.7 se presenta un ejemplo del funcionamiento del modelado espectral. El primer espectrograma muestra la señal ruidosa original con una SNR de 16 dB; el segundo exhibe las componentes transitorias estimadas; el tercero corresponde a las componentes armónicas obtenidas tras sustraer dichos transitorios; y el cuarto muestra el modelado sinusoidal aplicado sobre la parte armónica residual.

Como puede observarse en los espectrogramas segundo y tercero, los transitorios se separan correctamente de la parte armónica de la señal, aunque el ruido permanece presente en ambas representaciones, tal cual se muestra en la Ecuación 2.27. En cambio, en el último espectrograma —correspondiente al modelado sinusoidal— se aprecia que este enfoque no preserva el ruido en todo el espectro, lo cual permite aislar de forma precisa la componente sinusoidal de la señal, como lo denota la Ecuación 2.28.

A continuación, se define la señal

$$y^* = y|_v + x_s = (x_t + x_e|_v + n|_v) + x_s,$$

la cual se sustrae de la señal ruidosa y para obtener el residuo

$$r = y - y^* = x_e|_h + n|_h,$$

que mantiene las componentes estocásticas de la señal y el ruido resultantes tras el filtrado de mediana horizontal. Esta señal residual se utiliza como entrada para la segunda etapa del algoritmo: la sustracción espectral iterativa. Tal como se ilustra en la Figura 2.6, dicha etapa se divide en dos fases: (i) una sustracción iterativa aplicada sobre dicho residuo, preservando la señal $y*$ previamente calculada, y (ii) una sustracción iterativa aplicada a la señal completa.

Esta estrategia permite, en primera instancia, realizar una atenuación más intensa del ruido presente en la señal residual ($n|_h$), mediante un mayor número de iteraciones. Posteriormente, se aplica una sustracción más suave sobre la señal total, abarcando tanto la residual r como las componentes transitorias x_t, sinusoidales x_s y estocásticas x_e.

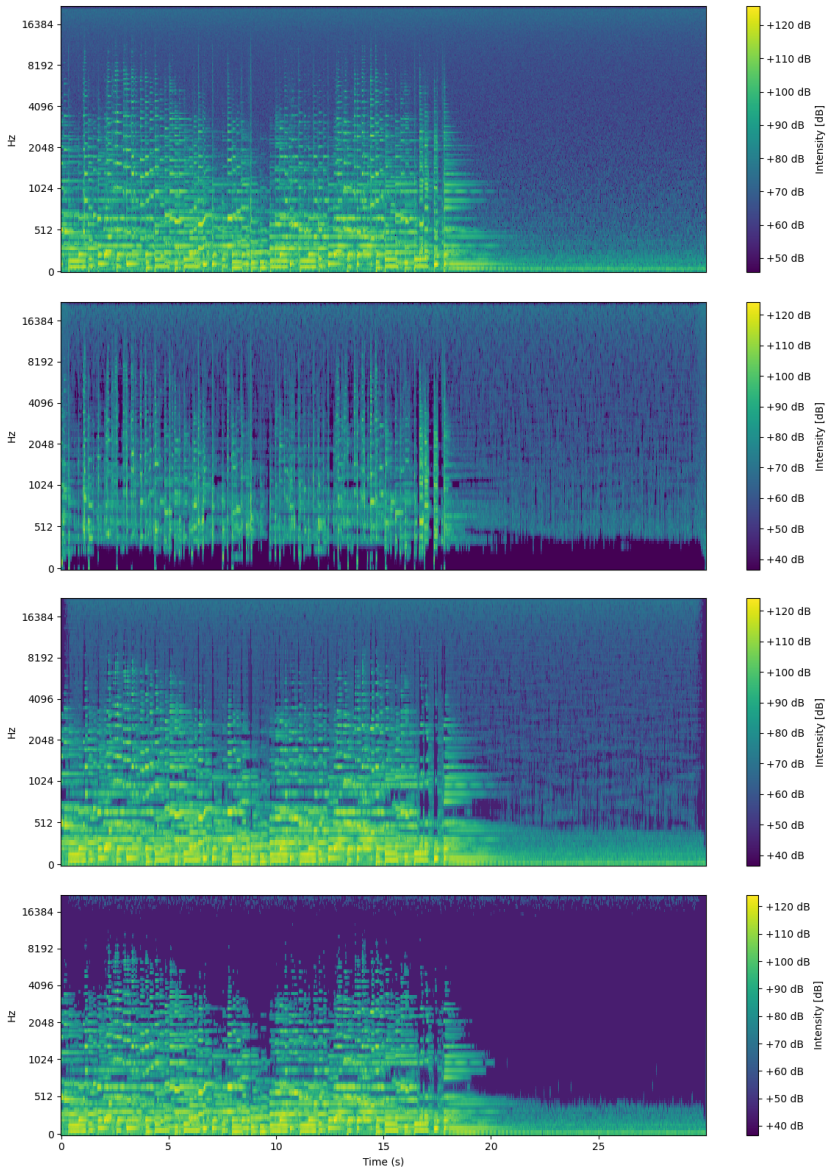

Figura 2.7: Ejemplo del proceso de modelado espectral. El primer espectrograma corresponde a la señal ruidosa original (SNR = 16 dB); el segundo muestra las componentes transitorias estimadas; el tercero presenta las componentes armónicas tras la sustracción de los transitorios; y el cuarto ilustra el modelado sinusoidal aplicado al residuo armónico.

En un escenario ideal, la primera fase de sustracción iterativa sobre r se encarga de atenuar las componentes del ruido asociadas al filtrado de mediana horizontal $n|_h$. A continuación, se reincorpora la señal y^* y se realiza la segunda fase de sustracción, destinada a reducir la componente restante $n|_v$. De este modo, se espera obtener una estimación de la señal limpia

$$\hat{x}^* \approx x_s + x_t + x_e = x.$$

Es importante destacar que, como se muestra en la Figura 2.6 y al igual que en el algoritmo básico descrito en la Sección 2.3, se emplean los *frames* inactivos, identificados mediante la detección de inactividad de la señal, para estimar el perfil de ruido que posteriormente se utilizará en la sustracción espectral de cada iteración, como se describió en algoritmo iterativo de la Subsección 2.5.1.

Finalmente, la señal obtenida tras la etapa iterativa del algoritmo $\hat{x}*$ se utiliza como entrada del método de reducción de ruido musical descrito en la Subsección 2.5.1, con el fin de aplicar un procesamiento final que atenúe este tipo de ruido en el resultado global del algoritmo. Esto resulta en la restauración final $\hat{x}$ de la señal x.

2.7. Parámetros de los algoritmos

En las Tablas 2.1–2.4 se resumen los parámetros utilizados por los algoritmos propuestos, organizados en función de su finalidad dentro del procesamiento. La Tabla 2.1 presenta los parámetros generales y aquellos asociados a la sustracción espectral, tanto en su versión clásica como iterativa. La Tabla 2.2 agrupa los parámetros empleados para la detección de inactividad o silencio en la señal de entrada. La Tabla 2.3 reúne los parámetros vinculados al modelado espectral, incluyendo tanto el análisis sinusoidal como la detección de transitorios. La Tabla 2.4 detalla los parámetros específicos para la reducción de ruido musical.

Tabla 2.1: Parámetros generales y de la Sustracción espectral (clásica e iterativa).

Parámetro	**Tipo**	**Descripción**
`x`	`ndarray`	Señal de entrada con ruido.
`fs`	`int`	Frecuencia de muestreo (Hz).
`nfft`	`int`	Tamaño de la FFT para el análisis y síntesis STFT.
`hop`	`int`	Tamaño del salto para el análisis y síntesis STFT.
`alpha`	`float`	Factor de sobre-sustracción.
`beta`	`float`	Factor de suelo espectral.
`rho`	`float`	Factor de suavizado del filtro paso bajo.
`n_iter`	`int`	Número de iteraciones de la sustracción espectral.
`sm_keep_pct`	`float`	Porcentaje de iteraciones en las que se preserva el modelo espectral (0–1).

Tabla 2.2: Parámetros de la Detección de inactividad de la señal.

Parámetro	Tipo	Descripción
th_energy	float	Umbral de energía para la detección de silencio.
th_zcr	float	Umbral de tasa de cruces por cero para la detección de silencio.
th_he	float	Umbral de magnitud en alta frecuencia para la detección de silencio.
zcr_hf_pct_cut	float	Porcentaje de corte en alta frecuencia para el cálculo de ZCR.
min_silence_len	int	Duración mínima (en *frames*) de un segmento de silencio.
min_sound_len	int	Duración mínima (en *frames*) de un segmento sonoro.
start_silence	int	Número mínimo de *frames* de silencio al inicio.
end_silence	int	Número mínimo de *frames* de silencio al final.
num_init_frames	int	Número de *frames* iniciales para referencia de ruido.

Tabla 2.3: Parámetros del Modelado espectral.

Parámetro	Tipo	Descripción
sm_nfft	int	Tamaño de la FFT para el modelado sinusoidal.
sm_hop	int	Tamaño del salto para el modelado sinusoidal.
peak_thresh	float	Umbral para detección de picos (dB).
min_sine_dur	float	Duración mínima de una sinusoide (segundos).
max_sines	int	Número máximo de sinusoides por *frame*.
fdev_offset	float	Desplazamiento de desviación de frecuencia para la continuación de sinusoides.
fdev_slope	float	Pendiente de desviación de frecuencia para la continuación de sinusoides.
td_nfft	int	Tamaño de la FFT para detección de transitorios.
td_Lh	int	Longitud del filtro mediano horizontal (en segundos o *frames*).
td_Lp	int	Longitud del filtro mediano percusivo (en Hz o bins).

Tabla 2.4: Parámetros de la Reducción de ruido musical.

Parámetro	Tipo	Descripción
mn_nfft	int	Tamaño de la FFT para reducción de ruido musical.
mn_hop	int	Tamaño del salto para reducción de ruido musical.
mn_thresh_db	float	Umbral (en dB) para supresión de ruido musical.
mn_win_len	int	Longitud de la ventana de suavizado para eliminación de ruido musical.

Capítulo 3

Aprendizaje profundo

En este capítulo se presentan las técnicas de aprendizaje profundo empleadas para la reducción de ruido en grabaciones musicales. En primer lugar, se revisan brevemente los antecedentes más relevantes del uso de redes neuronales en procesamiento de audio, con especial énfasis en su aplicación al *denoising*. A continuación, se describe en detalle el modelo en dos etapas propuesto por Moliner et al. [6], que constituye la base de este trabajo.

3.1. Introducción

El uso del aprendizaje automático para resolver problemas de procesamiento de señales ha crecido de forma significativa en los últimos años, y el procesamiento de música no es la excepción. El trabajo [32], muestra cómo en la última década los artículos que aplican aprendizaje profundo en música pasaron de poco más de diez en 2014 a más de doscientos en 2021.

El uso de aprendizaje automático permite superar las limitaciones de los métodos clásicos, los cuales suponen características del ruido (como estacionareidad o su distribución espectral) que no se cumplen estrictamente en contextos reales, teniendo que tratar las imperfecciones con técnicas independientes.

En 2020, Li et al. presentaron en [33] un algoritmo de *denoising* supervisado orientado a la restauración de grabaciones musicales históricas. Con los avances en aprendizaje profundo, los métodos basados en datos ofrecieron una alternativa flexible: eliminar del procesamiento los métodos tradicionales, no imponiendo suposiciones explícitas sobre el ruido y aprendiendo directamente de datos reales.

Este enfoque introduce nuevos desafíos: por un lado, diseñar un modelo capaz de capturar la complejidad estructural de la música. manteniendo una arquitectura suficientemente simple para ser entrenable, y por otro, construir un conjunto de datos adecuado, ya que las grabaciones musicales antiguas o degradadas carecen de versiones limpias de referencia [33].

El modelo propuesto por Li convierte internamente la señal en su representación tiempo–frecuencia mediante la STFT. El espectrograma resultante (representado como una imagen de dos canales, correspondientes a las componentes real e

imaginaria) se procesa con una red neuronal convolucional tipo *U-Net* 2D. Finalmente, la señal se reconstruye en el dominio temporal aplicando la iSTFT. Este proceso se puede observar en la Figura 3.1.

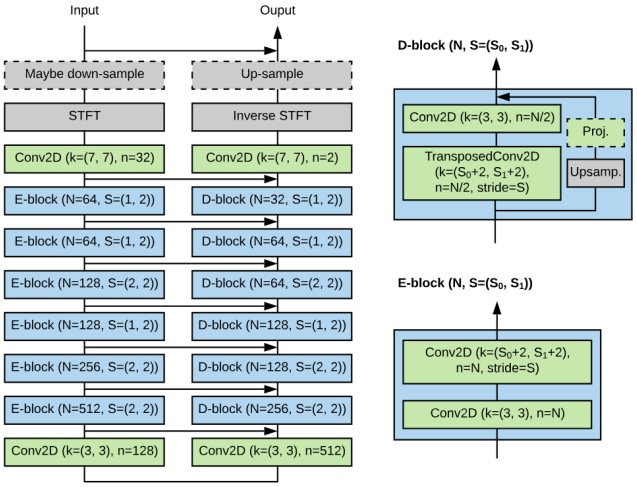

Figura 3.1: Esquema del método propuesto por Li et al. [33] para la restauración de grabaciones musicales históricas. El modelo convierte la señal de audio al dominio tiempo–frecuencia mediante la STFT, procesa el espectrograma complejo con una arquitectura *U-Net* 2D y reconstruye la señal en el dominio temporal mediante la iSTFT. Imagen tomada de [33].

El trabajo de Li fue continuado en 2022 por Moliner et al. [6], quienes realizaron diversas modificaciones a la arquitectura propuesta en [33] y entrenaron la red para el caso de ruido de gramófono. De esta manera, se propuso, por un lado, el uso de datos de ruido más realistas y, por otro, una arquitectura refinada basada en dos etapas de *U-Net*.

Los autores comentan que, en distintas pruebas con música artificialmente contaminada con ruido, el sistema alcanzó una calidad perceptual indistinguible del audio limpio original según la evaluación subjetiva de oyentes, siendo capaz de eliminar *colored noise*, *rumble* y eventos impulsivos [6]. En las secciones siguientes se describirá en detalle el trabajo desarrollado por Eloi Moliner et al., sobre el cual se basa este estudio, dado que su implementación y documentación son de libre acceso.

Posteriormente, este enfoque fue retomado en 2023 por Irigaray et al. [4], quienes aplicaron y adaptaron la metodología al problema de la reducción de ruido en grabaciones analógicas en cinta magnética. En este contexto, los autores destacan —en consonancia con Moliner et al. [6]— que uno de los factores decisivos para el alto rendimiento obtenido fue el uso de datos de ruido realistas. Para ello, desarrollaron una base de datos específica de ruido de cinta magnética [34], registrada a partir de diversos equipos funcionales. Con dicho material, entrenaron el modelo de aprendizaje automático propuesto en [6], empleando mezclas entre estos ruidos y fragmentos musicales limpios bajo distintos niveles de SNR.

Además, tanto la evaluación objetiva como la subjetiva confirmaron la eficacia

del método en la restauración de grabaciones analógicas, resaltando nuevamente los beneficios de entrenar con ruido real proveniente del dominio de aplicación específico.

3.2. Modelo de dos etapas *U-Net*

En esta sección se presenta el trabajo de *Eloi Moliner* y *Vesa Välimäki*, titulado "A two-stage U-Net for high-fidelity denoising of historical recordings" [6], orientado a la reducción del ruido en grabaciones históricas. En primer lugar, se describe el preprocesamiento aplicado a los datos de entrenamiento y, posteriormente, se detalla la arquitectura propuesta e implementada por los autores.

3.2.1. Preprocesamiento de los datos

En términos generales, los datos ruidosos utilizados durante el entrenamiento se crearon según la siguiente expresión:

$$X = \beta\big(Y + \alpha N\big), \tag{3.1}$$

donde X denota la señal contaminada, Y la señal limpia, N el ruido, α un factor de escalado que determina la SNR resultante y β un factor que ajusta el nivel global de la mezcla. La variación de estos parámetros introduce diversidad en las condiciones de entrenamiento, aumentando la robustez del modelo frente a distintos niveles de ruido e intensidad al simular múltiples escenarios de grabación y degradación. Esta estrategia forma parte de la técnica conocida como *data augmentation*.

La base de datos de grabaciones limpias considerada fue *MusicNet* [35], cuya descripción se presenta en la Subsección 4.1.1. Para evitar sesgos asociados a artefactos no deseados, se descartaron las grabaciones más antiguas del conjunto, cuya calidad se encontraba sensiblemente deteriorada. En cuanto al conjunto de ruidos, los autores utilizaron fragmentos extraídos del proyecto *"The Great 78 Project"* [36], descrito en la Subsección 4.1.3.

En la Figura 3.2 se muestran los diagramas de bloques correspondientes a los procedimientos utilizados para generar los datos ruidosos de entrenamiento.

Para construir cada audio limpio, primero se barajan (*shuffle*) las grabaciones de la base *MusicNet* para aleatorizar su orden de acceso y procesamiento. Luego, cada señal se carga de forma individual, se convierte a mono (si corresponde) y se normaliza por su valor máximo absoluto. La señal resultante se divide en *frames* sin solapamiento y de longitud fija, rellenando con ceros (*zero-padding*) cuando su duración es menor que la requerida.

Para cada *frame*, se seleccionan aleatoriamente un valor de SNR en el rango de 2 a 20 dB y un valor de escalado entre –6 y 4 dB. Con estos parámetros y un cierto segmento de ruido, se ajusta el nivel de la señal ruidosa de modo de obtener la SNR y la escala especificadas según la Ecuación 3.1.

La generación de los audios de validación sigue esencialmente el mismo procedimiento utilizado para el conjunto de entrenamiento. La única diferencia significativa es que no se aplica el *shuffle* inicial a la lista de grabaciones, por lo que

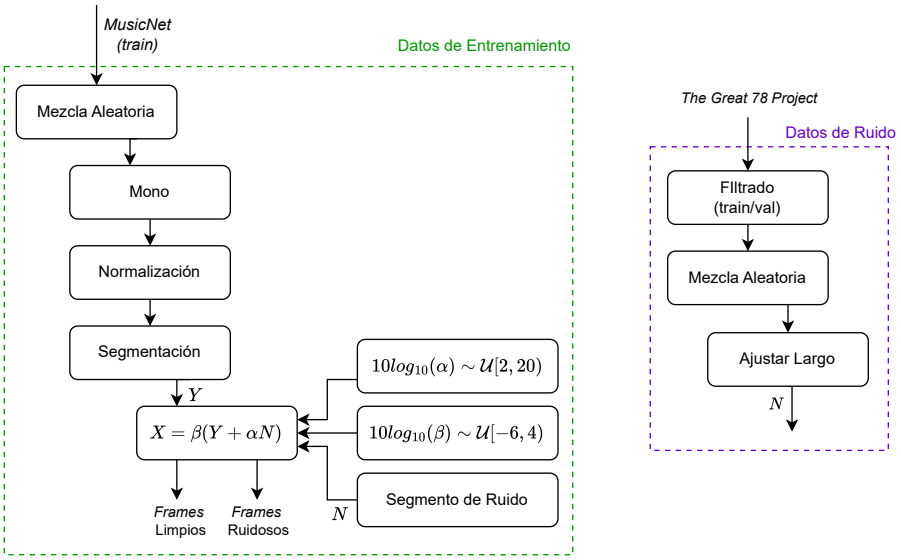

Figura 3.2: Diagrama de flujo del generador de datos de entrenamiento (izquierda) y del generador de segmentos de ruido (derecha). El primer bloque muestra las etapas de mezcla, conversión a mono, normalización, segmentación y generación de *frames* con SNR y nivel de escala aleatorios. El segundo bloque ilustra el proceso de selección y preparación de los segmentos de ruido.

las señales se procesan en el orden original en que aparecen en la base de datos destinada a validación.

Por otro lado, el preprocesamiento de los segmentos de ruido, también ilustrado en la Figura 3.2, sigue una serie de pasos específicos. En primer lugar, se filtran los audios de ruido correspondientes al conjunto de entrenamiento o validación, según sea necesario. A continuación, se realiza una mezcla aleatoria entre las señales filtradas. Finalmente, se ajusta la duración de cada segmento para que coincida con la longitud fija utilizada en los *frames* de los audios limpios.

Este ajuste consiste en recortar el segmento cuando su longitud excede la requerida, o bien extenderlo cuando resulta más corto. En este último caso, la extensión se realiza mediante un procedimiento *overlap–add* con ventanas de Hann: la señal se repite de forma circular, utilizando una periodicidad coherente con la rotación de discos de 78 rpm, y cada repetición se solapa suavemente con la anterior gracias a la ponderación de la ventana, evitando así discontinuidades audibles en los límites.

3.2.2. Descripción de la arquitectura

La arquitectura propuesta en [6] está compuesta por dos subredes *U-Net* conectadas en serie, con distintas entradas y objetivos de entrenamiento específicos, complementadas por un módulo de atención supervisada (*Supervised Attention Module*, SAM), como se ilustra en la Figura 3.3.

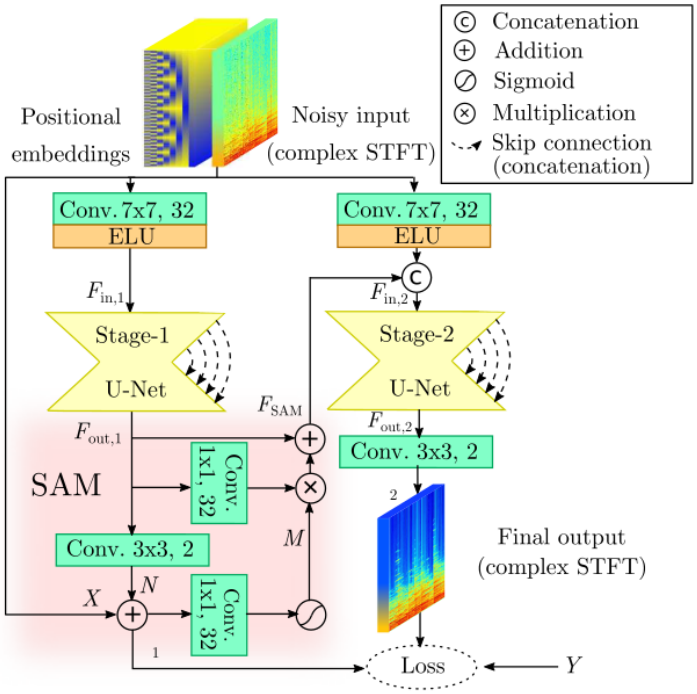

Figura 3.3: Arquitectura propuesta en [6], compuesta por dos subredes *U-Net* en serie y un módulo de atención supervisada (SAM). La primera *U-Net* modela el ruido residual, mientras que la segunda refina la estimación utilizando las representaciones generadas en la etapa previa. Imagen extraída de [6].

Esta separación permite abordar el *denoising* en dos fases sucesivas: la primera subred se encarga de estimar el ruido residual presente en la señal, mientras que la segunda realiza una atenuación refinada, considerando tanto la señal ruidosa como las estimaciones generadas por la primera etapa.

El modelo opera directamente sobre la STFT de la señal, utilizando como canales independientes las partes real e imaginaria del espectrograma. Para su cálculo se emplea una ventana de *Hamming* de 2048 muestras y un desplazamiento de 512 muestras.

A estos dos canales se le suman diez canales adicionales que corresponden a los llamados *frequency-positional embeddings*, los cuales permiten que las primeras capas convolucionales tomen en cuenta explícitamente la posición de cada componente espectral en el eje de frecuencia.

Cada uno de estos vectores dependen únicamente de la frecuencia f del *bin* correspondiente, y se construyen utilizando funciones coseno de diferentes frecuencias. La fórmula general es la siguiente:

$$\rho(f) = \left(\cos\left(\pi\frac{f}{F}\right), \cos\left(2\pi\frac{f}{F}\right), \ldots, \cos\left(2^{k-1}\pi\frac{f}{F}\right)\right), \tag{3.2}$$

donde F representa el ancho de banda total del espectrograma y $k = 10$ es el número de componentes del vector.

Esta técnica está inspirada en los *positional embeddings* utilizados en arquitecturas tipo *Transformer*, y tiene como objetivo proporcionar al modelo una representación explícita de la ubicación en frecuencia. Al concatenar estos vectores con los canales originales, el modelo puede distinguir mejor las características del espectro según su posición [37].

En cada etapa del modelo, la entrada de 12 canales se procesa inicialmente mediante un extractor de características compuesto por una capa convolucional seguida de una función de activación no lineal *Exponential Linear Unit* (ELU), tal como se ilustra en la Figura 3.3. En la primera etapa, las características obtenidas, denotadas como $F_{\text{in},1}$, se alimentan directamente a la subred *U-Net*. En cambio, en la segunda etapa, las características de entrada $F_{\text{in},2}$ se construyen concatenando las características generadas por el módulo SAM de la etapa anterior, representadas por F_{SAM}.

El espectrograma limpio final, denotado como $\hat{Y}_2$, se obtiene procesando las características de salida de la segunda subred *U-Net*, $F_{\text{out},2}$, mediante una última capa convolucional de tamaño 3×3.

Arquitectura *U-Net*

La arquitectura *U-Net* tiene una estructura simétrica en forma de "U", compuesta por una etapa de codificación (*encoder*), que reduce progresivamente la resolución para capturar el contexto global, y una etapa de decodificación (*decoder*), que recupera la resolución original mediante operaciones de *upsampling*. En este enfoque, las *skip connections* constituyen un elemento fundamental, ya que enlazan directamente las capas correspondientes del *encoder* y del *decoder*, lo que permite conservar detalles locales relevantes al mismo tiempo que se integra información contextual de mayor nivel.

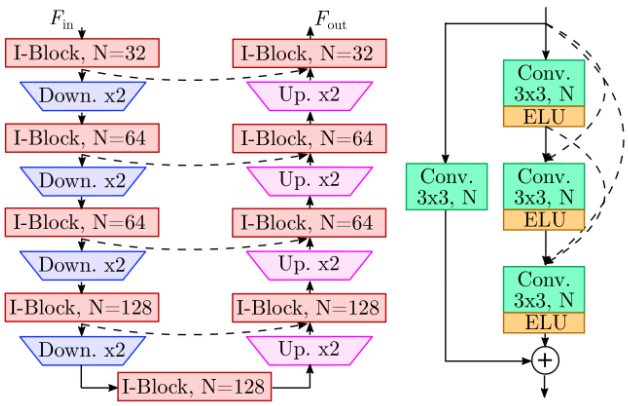

Figura 3.4: Estructura de la subred *U-Net*. La figura ilustra la estructura codificador-decodificador simétrica con cuatro niveles de reducción y expansión de resolución (izquierda), conectados mediante *skip connections*. Cada nivel incorpora un bloque intermedio denominado *I-Block* (derecha). El descenso en resolución se realiza mediante convoluciones con salto (*strided convolutions*). Imagen extraída de [6].

Tal como se muestra en la Figura 3.4, las *U-Net* empleadas en este estudio incorporan en cada nivel un bloque intermedio (*I-Block*) que actúa como un módulo de refinamiento local. Estos bloques combinan convoluciones densamente conectadas con una conexión residual, lo que permite capturar patrones espectrales finos sin perder estabilidad numérica ni capacidad de generalización. Su función es procesar la información a cada escala antes de modificar su resolución.

La ruta de codificación se implementa mediante convoluciones con salto (*strided convolutions*) de *kernel* 4×4 y paso 2×2. Estas operaciones no solo reducen la resolución temporal y frecuencial, sino que también expanden el campo receptivo de la red, permitiendo que los niveles más profundos integren información global del espectrograma. Esta característica es útil para modelar estructuras ruidosas amplias, como el *hiss* de banda ancha o patrones espectrales estables del ruido propio del soporte [6].

La etapa de decodificación reconstruye la resolución utilizando convoluciones transpuestas configuradas de manera simétrica respecto a las de la ruta de codificación. Durante esta fase, las características recuperan un mayor nivel de detalle y se combinan con las activaciones correspondientes del codificador mediante *skip connections*. Según los autores, estas conexiones evitan la pérdida de información local provocada por las operaciones de reducción de resolución y permiten que el modelo preserve bordes espectrales, armónicos débiles y transitorios relevantes que de otro modo podrían degradarse.

Módulo SAM

El módulo *SAM* se incorpora para ayudar al modelo a concentrarse en las regiones del espectrograma donde el ruido es más notorio. Su función es guiar la segunda etapa del proceso de *denoising* mediante un mecanismo de atención supervisada.

Durante el entrenamiento, este módulo aprende a generar un mapa que resalta las zonas tiempo-frecuencia con mayor presencia de ruido. Al aplicar este mapa sobre las representaciones internas del modelo, se refuerza la información relevante y se atenúa la menos útil, permitiendo una supresión de ruido más precisa en la segunda etapa de la red.

Como se puede notar en la Figura 3.3, el residuo estimado N se obtiene a partir de las características de salida de la primera subred *U-Net*, $F_{\text{out},1}$, mediante una capa convolucional de tamaño 3×3. La salida intermedia de la primera etapa, $\hat{Y}_1$, se calcula entonces como $\hat{Y}_1 = X + \hat{N}$, donde X es el espectrograma de entrada. A partir de esta salida, se generan las características F_{SAM} siguiendo el esquema ilustrado en dicha figura, utilizando máscaras de atención M que se calculan directamente a partir de $\hat{Y}_1$ mediante una convolución 1×1 seguida de una función sigmoide.

Parámetros del entrenamiento

Para entrenar el modelo, se empleó una función de pérdida basada en el error absoluto medio (*Mean Absolute Error*, MAE) entre las salidas de cada etapa y

el espectrograma limpio de referencia. La expresión matemática de la función de pérdida es:

$$\mathcal{L} = \frac{1}{K} \sum_{k} \left(\left| \hat{Y}_1^k - Y^k \right| + \left| \hat{Y}_2^k - Y^k \right| \right), \tag{3.3}$$

donde Y representa el espectrograma limpio y K la cantidad total de coeficientes de la STFT.

El proceso de entrenamiento se ejecuta de forma distribuida mediante la estrategia `MirroredStrategy` de `TensorBoard`, lo que permite aprovechar múltiples GPUs. Según los autores, el entrenamiento se llevó a cabo durante 300,000 *steps* (2000 por época) utilizando un tamaño de lote de 8 y un optimizador *Adam*. La tasa de aprendizaje inició en 1×10^{-4} y se redujo en un factor de 10 cada 100,000 *steps* (50 épocas). Además, no se aplicaron técnicas de normalización como *batch normalization* o *weight normalization*, ya que no mostraron mejoras en el rendimiento durante las pruebas [6].

Capítulo 4

Metodología

En el presente capítulo se describen en detalle los procedimientos metodológicos seguidos durante el desarrollo de este trabajo. En primer lugar, se definen las bases de datos utilizadas y las métricas consideradas para la experimentación. A continuación, se presenta el proceso de búsqueda de hiperparámetros llevado a cabo para los diferentes módulos de la implementación de sustracción espectral desarrollada en la sección 2.6. Posteriormente, se describe la metodología empleada para los entrenamientos del modelo propuesto en el Capítulo 3 y, finalmente, se detallan las estrategias de evaluación aplicadas a los modelos finales con el objetivo de analizar su desempeño en la reducción de ruido.

4.1. Bases de datos

4.1.1. Música clásica *(MusicNet)*

La base de datos *MusicNet*, introducida por John Thickstun, Zaid Harchaoui y Sham M. Kakade en el artículo *Learning Features of Music from Scratch* [35], consiste en 330 grabaciones de música clásica con licencia libre. Cada grabación incluye tanto el archivo de audio en formato WAV como anotaciones temporales detalladas de las notas musicales, instrumentos y compositores, almacenadas en un archivo CSV asociado. Los audios presentan una duración variable, se encuentran en formato mono, con una frecuencia de muestreo de 44.1 kHz y una resolución de 32 bits por muestra.

El conjunto contiene interpretaciones de obras de reconocidos compositores como Schubert y Mozart, entre otros. Algunos ejemplos incluidos son la *Piano Sonata in D major* de Schubert —con movimientos como *Allegro vivace* y *Scherzo. Allegro vivace*— y el *String Quartet No. 19 in C major* de Mozart —con movimientos como *Adagio–Allegro* y *Andante cantabile*.

4.1.2. Base de música personalizada

En el marco de este proyecto, se desarrolló una base de datos propia con el propósito de evaluar el desempeño de las técnicas de procesamiento propuestas en escenarios acústicos diversos. La colección incluye 48 grabaciones en formato estéreo, almacenadas como archivos WAV codificados a 16 bits por muestra y con una frecuencia de muestreo de 44,1 kHz. Las piezas abarcan distintos géneros musicales, épocas e instrumentaciones, y se organizan en cuatro categorías de 12 grabaciones cada una: Música Popular, Muchas Fuentes, Pocas Fuentes y Vocal.

La primera categoría **Música Popular** está conformada por grabaciones de géneros como rock, pop, blues, bossa nova y cumbia. Estas piezas presentan formaciones instrumentales típicas —guitarra eléctrica o acústica, bajo, batería y voz principal—, a menudo complementadas por teclados, percusión menor u otros instrumentos de acompañamiento. Este conjunto resulta representativo de producciones musicales modernas, con estructuras rítmicas marcadas y un equilibrio sonoro característico de mezclas comerciales. Algunos ejemplos incluidos son canciones de *Red Hot Chili Peppers*, *Jaime Roos* y *Madonna*, entre otros.

Luego, la segunda categoría **Muchas Fuentes** agrupa música con una alta densidad instrumental, que abarca desde orquestas clásicas hasta conjuntos contemporáneos con instrumentación diversa. Estas grabaciones se caracterizan por la coexistencia de numerosas fuentes —cuerdas, vientos, metales y percusión— que generan una elevada complejidad espectral y temporal. Entre los ejemplos se encuentran piezas de la banda sonora de *Indiana Jones* y *El Señor de los Anillos*, así como obras orquestales como *Liszt – Hungarian Rhapsody No. 2 in D minor, S.359 No. 2* interpretada por la *Orchestre symphonique de Montréal*, y *Pixinguinha e Sua Orquestra – Marreco Quer Água*.

La categoría **Pocas Fuentes** está compuesta por obras con un número reducido de instrumentos, tales como duetos de voz y guitarra o interpretaciones solistas. Este tipo de material permite analizar con mayor precisión los efectos del procesamiento sobre señales simples y bien definidas. Algunos ejemplos son baladas como *Blowin' in the Wind* (Bob Dylan), *Into My Arms* (Nick Cave) y *Someone Like You* (Adele).

Finalmente, la categoría **Vocal** reúne grabaciones a capela, incluyendo coros, interpretaciones de ópera sin acompañamiento y piezas contemporáneas centradas exclusivamente en la voz humana. Este conjunto resulta de particular interés, dado el papel fundamental de la voz en la mayoría de los contextos musicales y su relevancia para el estudio de la reducción de ruido en señales vocales. Ejemplos de esta categoría incluyen *Fernando Cabrera – Te Abracé en la Noche* y *Perotá Chingó – Coral*.

4.1.3. Ruido de gramófono

El conjunto de datos de gramófono (*gramophone record noise dataset*), desarrollado por Eloi Moliner y Vesa Välimäki [6], fue creado con el propósito de disponer de muestras de ruido altamente realistas para el entrenamiento de modelos de *denoising*. Para ello, se extrajeron segmentos de ruido a partir de grabaciones de

discos de gramófono de 78 rpm, pertenecientes a la colección pública y digitalizada del *The Great 78 Project* [36].

Las muestras incluyen una combinación de degradaciones procedentes de diversas fuentes: ruido eléctrico de los circuitos (como *hiss*), ruido ambiental del entorno de grabación, ruido de baja frecuencia (*rumble*) generado por el giradiscos, e irregularidades del soporte físico que producen clics y golpes (*clicks* y *thumps*). Se profundiza sobre estas degradaciones en el Anexo B.1.

Para seleccionar automáticamente los segmentos que contenían únicamente ruido, Moliner y Välimäki entrenaron un clasificador binario basado en redes neuronales, con una arquitectura similar a la propuesta en PoCoNet por İsik et al. [37], utilizando un subconjunto de ejemplos etiquetados manualmente como referencia. Este enfoque permitió reducir los falsos positivos habituales en métodos basados únicamente en umbrales de energía, los cuales tienden a confundir pasajes musicales suaves, colas de reverberación o desvanecimientos con ruido puro [6, 33].

El conjunto final comprende 139 minutos de audio en mono, con una resolución de 16 bits por muestra y una frecuencia de muestreo de 44,1 kHz, divididos en 2430 segmentos extraídos de 1386 grabaciones diferentes, entre los años 1902 y 1966.

4.1.4. Grabaciones analógicas de cintas de audio

El conjunto de datos de grabaciones en cinta de audio analógica (*Analog Audio Tape Recordings*) fue desarrollado por Ignacio Irigaray, Martín Rocamora y Luiz W. P. Biscainho [4], con el objetivo de caracterizar el ruido inherente al medio magnético y al mecanismo de reproducción. Para ello, se reprodujeron cintas vírgenes en distintos equipos de cinta abierta y cassette, registrando exclusivamente el ruido generado por el sistema sin contenido musical. Las grabaciones se realizaron en el Centro Nacional de Documentación Musical (Montevideo, Uruguay) utilizando una interfaz *M-Audio Fast Track Pro*, con una frecuencia de muestreo de 44,1 kHz y una resolución de 16 bits por muestra. Todos los equipos fueron previamente calibrados y mantenidos para garantizar su correcto funcionamiento.

Se utilizaron cinco grabadores de cinta abierta: dos modelos semi-profesionales Revox A77 (versiones *normal-speed* y *high-speed*), un grabador a válvulas Revox C-36, y dos grabadores portátiles Uher (modelos 4000 Report S y 4000 Report L). Además, se empleó un reproductor de cassette Technics TR-575 de doble deck, con un cassette virgen TDK HX-S60. Para las grabaciones en cinta abierta se utilizó *Premium Analog Recording Tape* de ATR Magnetics. Las sesiones abarcaron distintas velocidades de reproducción (*inches per second*, IPS) según el dispositivo: 1.875, 3.75, 7.5 y 15 IPS. En la Figura 4.1 se pueden apreciar algunos de los grabadores y reproductores utilizados para la creación de la base de datos.

El ruido característico del medio incluye principalmente *hiss* (ruido de alta frecuencia generado por la aleatoriedad del grano magnético y el ancho de banda del sistema), *hum* y *buzz* (componentes tonales producidas por interferencias eléctricas, típicamente en 50/60 Hz y sus armónicos), y ruido de modulación (variaciones del nivel de ruido dependientes de la señal grabada). En la Sección B.1 del Anexo B se describe cada una detalladamente.

Revox A77 | Uher 4000 Report L | Technics TR-575 | Revox C-36

Figura 4.1: Grabadores y reproductor utilizados en las sesiones de grabación analógica.

En total, el conjunto de datos contiene aproximadamente 2 horas de audio en mono, equivalentes a 10 minutos por cada combinación de dispositivo y velocidad.

4.1.5. MagTapeDB: Una base de datos de grabaciones históricas en cinta magnética

MagTapeDB [34] es una base de datos diseñada para el desarrollo y evaluación de técnicas de restauración de audio aplicadas a grabaciones musicales históricas almacenadas en cintas magnéticas. Su objetivo principal es proporcionar material realista que refleje las características y degradaciones propias del medio analógico, tales como *hiss*, *hum*, *wow and flutter*, saturación y caídas de señal, las cuales no suelen estar representadas en los conjuntos de datos sintéticos o modernos empleados habitualmente en el procesamiento digital de audio.

La colección está compuesta por más de 800 fragmentos de audio provenientes del archivo musicológico de Lauro Ayestarán [3]. En total, la versión actual del conjunto incluye 894 fragmentos (aproximadamente 351 minutos) distribuidos en tres categorías: grabaciones musicales, tonos de diapasón (*pitchpipe tones*) y segmentos de ruido de cinta. Cada archivo de audio cuenta con metadatos asociados que incluyen información como el número de carrete, año de grabación, velocidad de cinta, presencia de instrumentos, localización geográfica y, cuando es posible, frecuencia de afinación estimada.

Las digitalizaciones se realizaron a partir de cintas de 1/4 de pulgada en formato mono, reproducidas mediante una grabadora Revox A77 y digitalizadas a través de una interfaz *Universal Audio Apollo Solo*. Posteriormente, los audios fueron segmentados y anotados manualmente, identificando regiones musicales, tonos de referencia y fragmentos de ruido, a partir de los cuales se generaron extractos estandarizados de 30 segundos.

4.2. Métricas para la evaluación

En esta sección se describen las métricas utilizadas para evaluar los modelos finales obtenidos. Estas métricas permiten cuantificar de forma objetiva el desempeño de los distintos enfoques analizados y establecer una base de comparación consistente entre los resultados experimentales presentados posteriormente.

4.2.1. Error Cuadrático Medio Relativo (RMSE)

El *Error Cuadrático Medio Relativo*, o *Relative Mean Square Error* (RMSE), es una métrica utilizada para cuantificar el error promedio entre una señal estimada y su referencia, normalizado con respecto a la energía de la señal original. De esta manera, se puede expresar la magnitud del error en términos relativos, facilitando la comparación entre señales de distinta escala o amplitud. Matemáticamente, se define como:

$$\text{RMSE} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} y_i^2} \tag{4.1}$$

donde y_i representa la señal de referencia, $\hat{y}_i$ la señal estimada y N el número total de muestras.

El valor de RMSE es adimensional y toma valores no negativos. Un RMSE igual a cero indica una coincidencia perfecta entre ambas señales, mientras que valores mayores reflejan un incremento proporcional del error relativo.

4.2.2. Precisión, Recuperación y F_β-Score

Las métricas de *Precisión* (*Precision*) y *Recuperación* (*Recall*) son ampliamente utilizadas en problemas de clasificación y detección, ya que permiten evaluar el desempeño de un sistema en términos de su capacidad para identificar correctamente los elementos de interés.

La **Precisión** mide la proporción de verdaderos positivos entre todas las predicciones positivas realizadas por el sistema, y se define como

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

donde TP corresponde a los *verdaderos positivos* (instancias correctamente detectadas) y FP a los *falsos positivos* (instancias incorrectamente clasificadas como positivas). Una alta precisión indica que la mayoría de las detecciones son correctas, es decir, que el sistema comete pocos falsos positivos.

Por otro lado, la **Recuperación** o **Sensibilidad** (*Recall*) cuantifica la proporción de verdaderos positivos detectados respecto al total de positivos reales, y se expresa como

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

donde FN representa los *falsos negativos* (instancias positivas que el sistema no logró detectar). Un alto valor de *Recall* implica que el sistema detecta la mayoría de los casos relevantes, aunque podría incluir algunos errores adicionales.

Dado que la Precisión y el Recall tienden a presentar un compromiso entre sí, se introduce el **F_β-Score**, una medida combinada que pondera ambas métricas según un parámetro β que controla la importancia relativa de la Recuperación frente a la Precisión. Su definición general es

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \tag{4.4}$$

Cuando $\beta = 1$, ambas métricas se ponderan de forma equitativa, obteniéndose el conocido F_1-*Score*. Valores de $\beta > 1$ otorgan mayor peso a la Recuperación, mientras que valores de $\beta < 1$ favorecen la Precisión.

4.2.3. Evaluación Perceptual de la Calidad del Audio (PEAQ)

La *Evaluación Perceptual de la Calidad del Audio*, o *Perceptual Evaluation of Audio Quality* (PEAQ), desarrollada bajo la Recomendación *ITU-R BS.1387* [38], constituye una de las métricas más reconocidas y utilizadas en la industria del audio para la evaluación objetiva de la calidad percibida. Este método emplea un modelo psicoacústico que simula el funcionamiento del sistema auditivo humano con el fin de identificar distorsiones y artefactos introducidos en señales procesadas, generando un valor objetivo que se correlaciona estrechamente con las evaluaciones subjetivas realizadas por oyentes expertos.

En este trabajo se emplea la implementación *GstPEAQ*, disponible públicamente en el repositorio de *GitHub* [39]. Dicha herramienta reproduce el algoritmo descrito en la Recomendación *ITU-R BS.1387-1* [38], y permite realizar mediciones objetivas de la calidad percibida del audio tanto en su versión básica como avanzada.

La degradación perceptual de una grabación se cuantifica mediante la **Calificación de Diferencia Objetiva** (*Objective Difference Grade*, ODG), la cual busca aproximar la puntuación promedio que otorgaría un oyente humano experto. Este índice se expresa en una escala continua cuyos valores de referencia se presentan a continuación:

- **0**: Degradación imperceptible.
- **-1**: Degradación perceptible, pero no molesta.
- **-2**: Degradación levemente molesta.
- **-3**: Degradación molesta.
- **-4**: Degradación muy molesta.

4.2.4. Medida Perceptual de la Calidad del Audio (PAQM)

De forma similar a PEAQ, la *Medida Perceptual de la Calidad del Audio*, o *Perceptual Audio Quality Measure* (PAQM), es una métrica desarrollada para evaluar de forma objetiva la calidad percibida de señales de audio, basada en un modelo psicoacústico que simula el comportamiento del sistema auditivo humano. Este enfoque permite estimar el grado de degradación introducido por un procesamiento o codificación al comparar una señal procesada con su versión original.

En este trabajo se emplea la implementación de PAQM disponible en el repositorio de *GitHub* [40], desarrollada en *PyTorch* por *J. G. Beerends* y *J. A. Stemerdink*, según lo descrito en su publicación original [41].

El modelo produce un puntaje que varía a partir de cero, donde un valor de **0** indica que la señal procesada es indistinguible de la original, reflejando una degradación perceptual mínima o inexistente. A medida que el valor aumenta, se interpreta una mayor diferencia perceptual entre ambas señales.

Esta métrica se utiliza en este trabajo como referencia complementaria a PEAQ para la evaluación perceptual de la calidad de las grabaciones.

4.2.5. Relación señal–ruido estimada

Vale la pena recordar que la relación señal a ruido se define como

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{P_{\text{señal}}}{P_{\text{ruido}}} \right) \tag{4.5}$$

donde $P_{\text{señal}}$ y P_{ruido} representan las potencias promedio de la señal y del ruido, respectivamente.

En este trabajo se emplea una versión **estimada** de la SNR, calculada a partir de la representación espectro-temporal de la señal mediante la STFT. Para ello, se calcula la energía promedio de cada trama temporal y se distingue entre regiones con y sin señal de interés mediante un vector indicador.

La potencia promedio de cada trama se obtiene promediando sobre las bandas de frecuencia:

$$P[m] = \frac{1}{K} \sum_{k=1}^{K} |X(k,m)|^2, \tag{4.6}$$

donde $X(k,m)$ es el valor complejo de la STFT en la banda de frecuencia k y el *frame* temporal m, y K es el número total de bins de frecuencia. De esta manera, $P[m]$ representa la potencia promedio de la señal en el *frame* temporal m.

A partir de estas potencias, se definen los valores promedio en los *frames* de señal (p_{sr}) y de ruido (p_r), y la estimación final de la SNR se calcula como:

$$\mathrm{SNR}_{\text{estimado}} = 10 \log_{10} \left(\frac{p_{sr} - p_r}{p_r} \right). \tag{4.7}$$

Esta formulación no requiere disponer de una señal de referencia limpia, lo que la hace especialmente útil para evaluar grabaciones reales o procesadas, donde solo se tiene acceso a la señal resultante. De este modo, la SNR estimada proporciona una medida objetiva del predominio de la señal útil sobre el ruido residual, permitiendo cuantificar la calidad o la mejora alcanzada tras el procesamiento.

4.3. Búsqueda de hiperparámetros

La implementación final descrita en la Sección 2.6 presenta una complejidad considerable debido al elevado número de parámetros y configuraciones involucradas. Por esta razón, se optó por realizar una búsqueda de hiperparámetros de manera modular, abordando cada componente del sistema de forma independiente.

En primer lugar, se ajustaron los parámetros asociados al detector de inactividad, con el fin de establecer una segmentación adecuada entre regiones de silencio y de señal.

Posteriormente, se realizaron dos procesos de búsqueda diferenciados:

- **Búsqueda de hiperparámetros del algoritmo básico de sustracción espectral:** destinada a optimizar el rendimiento del método clásico, considerando parámetros como α, β y ρ.
- **Búsqueda de hiperparámetros de la implementación propuesta *SS Denoisify*:** enfocada en ajustar los parámetros de la versión iterativa desarrollada y del módulo de reducción de ruido musical.

Esta estrategia permitió aislar el efecto de cada conjunto de parámetros sobre el desempeño global del sistema, facilitando la comparación directa entre el algoritmo clásico y la propuesta desarrollada en este trabajo.

Vale la pena aclarar que en todos los casos, a excepción del último módulo, se utilizó una longitud de ventana FFT de `nfft` = 2048 muestras, con un desplazamiento (*hop size*) equivalente a una cuarta parte de dicha longitud (`hop` = `nfft` / 4).

4.3.1. Detector de inactividad

Para el ajuste de los parámetros del algoritmo de detección de inactividad en la señal, se seleccionaron aleatoriamente dos audios de cada grupo perteneciente a la base de datos de música personalizada. En cada uno de ellos se etiquetaron manualmente los segmentos temporales (en segundos) correspondientes a regiones de silencio, es decir, aquellos tramos que el algoritmo debería identificar como inactivos. Además, se incluyeron tres audios de ruido provenientes de la base de datos *Analog Audio Tape Recordings* (con una SNR de 10 dB), correspondientes a los dispositivos Revox A77, Uher 4000 Report S y Uher 4000 Report L. El análisis se realizó considerando los últimos 30 segundos de cada grabación.

A cada segmento de audio limpio se le asignó un segmento de ruido correspondiente a cada uno de los tres audios mencionados. Los tramos de ruido fueron seleccionados sin solapamiento, de modo de evitar correlaciones entre los distintos audios evaluados. En total, el procedimiento se aplicó sobre $8 \times 3 = 24$ señales (ocho audios limpios y tres tipos de ruido).

Dado un conjunto de configuraciones posibles para el algoritmo, el proceso de búsqueda consistió en evaluar cada combinación de parámetros sobre los audios seleccionados, con el objetivo de cuantificar la eficiencia con que el detector identifica los intervalos inactivos.

Para ello, las señales fueron segmentadas en *frames* de duración fija, y para cada trama se determinó si pertenecía a una región activa o inactiva, tanto en las etiquetas manuales como en la salida del algoritmo. De esta manera, cada configuración de parámetros produjo una secuencia binaria de detecciones, que fue comparada con la secuencia de referencia mediante la métrica $\mathbf{F}_\beta$**-Score**.

En este trabajo se utilizó un valor de $\beta = 0{,}8$, lo que otorga un mayor peso a la precisión que a la recuperación. De esta forma, el detector se penaliza más por clasificar erróneamente una trama activa como inactiva (FP) que por omitir una región silenciosa real (FN). Este criterio busca favorecer un comportamiento conservador del detector, privilegiando la fiabilidad en la identificación de los silencios reales.

Complementariamente, se buscó minimizar el **RMSE** entre el perfil de ruido calculado a partir de los segmentos de 30 segundos y el perfil detectado sobre el audio contaminado con ruido, de manera de asegurar una adecuada estimación del ruido de fondo. Para ello, ambos perfiles se ponderaron utilizando la curva de *A-weighting* (que se observa en la Figura 4.2), un filtro que refleja la sensibilidad del oído humano a distintas frecuencias, otorgando mayor relevancia a aquellas bandas donde la percepción auditiva es más sensible y reduciendo el peso de las frecuencias menos audibles [42]. Esta ponderación permite que la evaluación del ruido se alinee mejor con la percepción subjetiva de la calidad sonora.

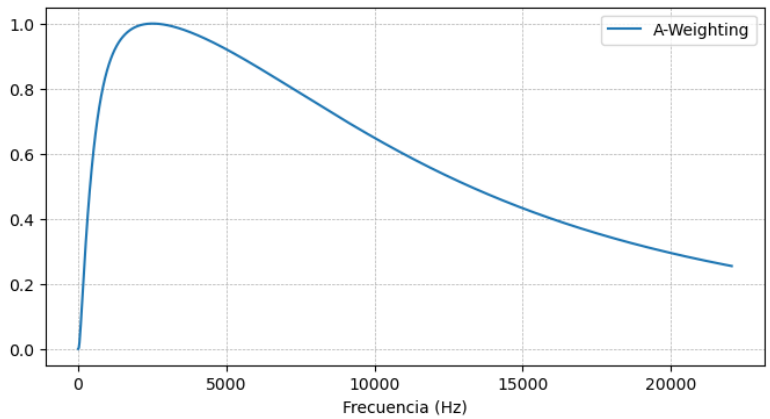

Figura 4.2: Curva de ponderación en A-weighting, mostrando cómo se ajustan los pesos de las distintas frecuencias para reflejar la sensibilidad del oído humano.

Tabla 4.1: Rangos de valores considerados en la búsqueda de hiperparámetros del detector de inactividad.

Hiperparámetro	**Valores evaluados**
`th_energy`	[0.15, 0.3, 0.45, 0.6, 0.75, 0.8, 0.85, 0.9, 0.95]
`th_zcr`	[0.15, 0.25, 0.3, 0.35, 0.45, 0.6, 0.75, 0.8, 0.86, 0.9, 0.92, 0.95]
`th_he`	[0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.45, 0.6, 0.75]
`zcr_hf_pct_cut`	[0.25, 0.45, 0.65, 0.85, 0.88, 0.9, 0.91, 0.94]
`min_silence_len`	[5, 10, 15, 20, 30]
`min_sound_len`	[10, 20, 25, 30, 35, 40]
`start_silence`	[1, 2, 4, 5, 6, 8]
`end_silence`	[1, 2, 4, 6, 8]
`num_init_frames`	[5]

Los rangos de valores considerados para los parámetros del algoritmo de detección de inactividad se presentan en la Tabla 4.1. Cada hiperparámetro fue evaluado sobre un conjunto de valores representativos, con el objetivo de identificar la combinación que optimiza la detección de los *frames* inactivos en la señal.

4.3.2. Sustracción espectral

Una vez obtenidos los resultados de la búsqueda de hiperparámetros para el detector de inactividad, se llevaron a cabo dos nuevas búsquedas: una correspondiente al algoritmo básico de sustracción espectral y otra asociada a la implementación propuesta en la Sección 2.6.

En ambos casos, se emplearon las mismas grabaciones utilizadas en la etapa anterior y se evaluó el desempeño de cada configuración mediante tres métricas principales: **PEAQ**, **PAQM** y la **SNR estimada**.

Para cada par de segmentos de audio limpio y ruido (a una SNR de 10 dB), se calculó inicialmente el deterioro perceptual entre la señal original y la señal ruidosa mediante las métricas PEAQ y PAQM. Posteriormente, la señal contaminada fue procesada con el algoritmo de sustracción espectral correspondiente, obteniéndose las métricas finales respecto a la señal original.

Las diferencias entre las métricas iniciales y finales permiten cuantificar la mejora perceptual introducida por el algoritmo, según las siguientes expresiones:

$$\Delta PAQM = PAQM_{\text{inicial}} - PAQM_{\text{final}} \tag{4.8}$$

$$\Delta PEAQ = PEAQ_{\text{final}} - PEAQ_{\text{inicial}} \tag{4.9}$$

donde los valores positivos de $\Delta PAQM$ y $\Delta PEAQ$ indican una mejora perceptual en la señal procesada.

Adicionalmente, se calculó la SNR estimada final utilizando los *frames* previamente etiquetados como activos o inactivos de forma manual. Un mayor valor de la SNR estimada indica que la potencia del ruido resulta relativamente menor en comparación con la potencia de la señal original, lo cual puede corresponder con una mejora en la calidad de la señal restaurada.

En el caso de la implementación del método de sustracción espectral fue necesario tener en cuenta dos consideraciones principales: la elevada cantidad de parámetros disponibles y el tiempo de ejecución del algoritmo. Este último resultó considerablemente mayor que el del método clásico, debido a su naturaleza iterativa y a la inclusión de etapas adicionales de procesamiento, como el modelado espectral.

Por esta razón, se decidió fijar los parámetros asociados al modelado espectral siguiendo las recomendaciones presentadas en la bibliografía [28,30], con excepción del parámetro `peak_thresh`, que determina el umbral de potencia a partir del cual se realiza la detección de picos utilizada posteriormente en la síntesis sinusoidal.

Los parámetros sometidos a evaluación correspondientes al módulo iterativo del algoritmo fueron: `alpha`, `beta`, `n_iter` y `sm_keep_pct`. En este caso, se adoptó un valor de `rho` igual a 0.01, considerado el menos restrictivo posible, con el objetivo de permitir una mayor flexibilidad en las iteraciones del proceso de sustracción

espectral. Asimismo, se seleccionaron valores de `alpha` y `beta` menos restrictivos. Este enfoque busca lograr una sustracción más suave —mediante valores relativamente bajos de `alpha`— y un umbral superior más permisivo, evitando una eliminación excesiva de la energía espectral. En las Tablas 4.2 y 4.3 se presentan los rangos de valores considerados para la evaluación de ambos algoritmos de sustracción espectral.

Tabla 4.2: Rangos de valores considerados en la búsqueda de hiperparámetros para el algoritmo SS Clásico.

Hiperparámetro	**Valores evaluados**
alpha	[0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.84, 0.88, 0.9, 0.92, 0.94, 0.96, 0.99]
beta	[0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6]
rho	[0.01, 0.05, 0.1, 0.15]

Tabla 4.3: Rangos de valores considerados en la búsqueda de hiperparámetros para el algoritmo SS Denoisify.

Hiperparámetro	**Valores evaluados**
alpha	[0.1, 0.15, 0.2, 0.25]
beta	[0.93, 0.96, 0.99]
n_iter	[10, 20, 30]
sm_keep_pct	[0.5, 0.7, 0.9]
peak_thresh	[-60, -50, -40]

Es importante destacar que en este trabajo se optó por utilizar valores de `alpha` < 1, a diferencia de lo planteado originalmente en la técnica clásica de sustracción espectral. El uso de factores de atenuación mayores (`alpha` > 1) puede resultar adecuado en aplicaciones de procesamiento de voz, donde la supresión agresiva del ruido residual no afecta de manera significativa la inteligibilidad. Sin embargo, se comprobó experimentalmente que, en señales musicales, este enfoque tiende a eliminar energía espectral relevante, incluyendo armónicos y matices tímbricos perceptualmente importantes. En consecuencia, la señal procesada puede presentar una pérdida de naturalidad y una modificación perceptible de su timbre.

4.3.3. Algoritmo de reducción de ruido musical

Finalmente, se realizó la búsqueda de hiperparámetros correspondiente al algoritmo de reducción de ruido musical, asociado al último módulo de la implementación detallada en la Sección 2.6.

De forma análoga a las evaluaciones previas, se utilizaron los mismos segmentos de audio y las métricas objetivas **PEAQ** y **PAQM**. En este caso, no se consideró el uso de la **SNR estimada**, dado que las componentes espectrales asociadas al ruido musical presentan una potencia relativamente baja, por lo que su impacto en el valor global de SNR resulta despreciable.

Las variaciones de PEAQ y PAQM se calcularon con respecto a la salida del módulo iterativo de sustracción espectral, en lugar de hacerlo sobre el audio contaminado. De esta manera, la comparación refleja específicamente la mejora atribuible a la etapa de reducción de ruido musical y su efecto sobre la calidad perceptual del resultado final.

En la Tabla 4.4 se presentan los rangos de valores considerados en la búsqueda de hiperparámetros para este módulo. En particular, se evaluaron distintos tamaños de ventana FFT y longitudes de ventana temporal expresadas en segundos, de modo que:

$$\texttt{mn\_win\_len} = \textbf{int}(\texttt{mn\_win\_len (s)} \times \texttt{sampling\_rate})$$

Tabla 4.4: Rangos de valores considerados en la búsqueda de hiperparámetros para el algoritmo de reducción de ruido musical.

Hiperparámetro	**Valores evaluados**
mn_nfft	[256, 512, 1024, 2048, 4096, 8192]
mn_thresh_db	[-5, -10, -15, -20, -25, -30]
mn_win_len (s)	[1e-3, 5e-3, 1e-2, 5e-2, 1e-1]

4.3.4. Elección de la configuración óptima

Dado que las métricas mencionadas en las secciones anteriores no comparten necesariamente los mismos rangos ni escalas de valores, se decidió normalizarlas en el intervalo $[0, 1]$ con el fin de garantizar una comparación equitativa entre las distintas configuraciones evaluadas. Esta normalización permitió, además, definir una **métrica ponderada** que integra de manera conjunta la información de las distintas medidas de desempeño, facilitando la selección de la configuración óptima de parámetros para cada módulo.

Por cada configuración de hiperparámetros se calcularon la **media** y la **desviación estándar** de las métricas correspondientes a las señales utilizadas en la evaluación.

Sea una métrica X asociada a un conjunto de valores medios $\{x_0, x_1, x_2, \dots\}$ y desviaciones estándar $\{\sigma_0, \sigma_1, \sigma_2, \dots\}$, uno por cada configuración evaluada. En primer lugar, con el fin de penalizar configuraciones que presenten una alta variabilidad, se aplicó un ajuste suave a los valores medios. En el caso de las métricas donde un valor mayor implica una mejora —tales como $\Delta PEAQ$, $\Delta PAQM$, *SNR estimada* y F_β-Score— se utilizó la siguiente penalización:

$$\tilde{x}_i = \frac{x_i}{1 + \frac{\sigma_i}{100}}, \qquad i = 0, 1, 2, \dots \tag{4.10}$$

Por otro lado, para métricas donde una menor media representa un mejor desempeño, como es el caso del *RMSE*, la penalización se aplicó de manera inversa, es decir, multiplicando por el factor correspondiente:

$$\tilde{x}_i = x_i \left(1 + \frac{\sigma_i}{100}\right), \qquad i = 0, 1, 2, \ldots \tag{4.11}$$

Posteriormente, todos los valores ajustados $\tilde{x}_i$ se normalizaron en el intervalo $[0, 1]$ mediante una normalización `min-max`:

$$\bar{x}_i = \frac{\tilde{x}_i - \text{mín}(\tilde{x})}{\text{máx}(\tilde{x}) - \text{mín}(\tilde{x})}, \qquad i = 0, 1, 2, \ldots \tag{4.12}$$

De esta forma, todas las métricas quedan expresadas en una escala común, lo que habilita su combinación en una única métrica ponderada destinada a seleccionar la configuración óptima de hiperparámetros. A continuación, se detalla la ponderación aplicada en cada uno de los módulos, dado que cada componente del sistema persigue objetivos específicos y, en consecuencia, requiere priorizar determinadas métricas por encima de otras.

En el caso del detector de inactividad, la métrica final se estableció como una combinación ponderada del F_β-Score con un peso del 80 % y del *RMSE* con un 20 %. Esta elección refleja la importancia central del F_β-Score para evaluar la capacidad del detector de distinguir entre actividad e inactividad, mientras que el *RMSE*, aunque menos relevante para esta tarea, se incorpora para favorecer configuraciones que, además, permitan obtener estimaciones más representativas del ruido.

Para los algoritmos de sustracción espectral —tanto el método Clásico como la variante *SS Denoisify*— se definió una métrica final basada en una ponderación del 45 % para cada una de las métricas diferenciales $\Delta PEAQ$ y $\Delta PAQM$, y del 10 % para la SNR estimada. Esta asignación prioriza explícitamente el impacto perceptual del proceso de *denoising*, ya que las métricas diferenciales capturan de forma directa las mejoras o degradaciones respecto a la señal limpia, mientras que la SNR estimada, si bien aporta una medida complementaria, no siempre se correlaciona de manera consistente con la calidad perceptual.

Finalmente, en la búsqueda de hiperparámetros del algoritmo de reducción de ruido musical, también se emplearon las métricas diferenciales $\Delta PAQM$ y $\Delta PEAQ$, aunque en este caso calculadas con respecto a la salida del módulo iterativo de *SS Denoisify*. Como se mencionó anteriormente, esta elección permite evaluar específicamente si este último módulo aporta una contribución adicional al proceso de reducción del ruido residual. Dado que ambas métricas reflejan la calidad perceptual del resultado generado por el módulo, la métrica final se definió mediante una ponderación equitativa del 50 % para cada una.

4.4. Entrenamiento del modelo de aprendizaje profundo

Uno de los objetivos de este trabajo es presentar y comparar dos enfoques sustancialmente distintos para la restauración de grabaciones: la sustracción espectral y el aprendizaje automático. En las secciones anteriores se describió la metodología utilizada para optimizar el desempeño de ambos algoritmos —tanto el clásico como el propuesto— mediante la exploración sistemática de sus hiperparámetros.

En lo que respecta al aprendizaje profundo, en el Capítulo 3 se desarrolló un modelo específico basado en el trabajo de *Eloi Moliner* y *Vesa Välimäki* [6]. En la presente sección se detalla la metodología empleada para realizar distintos entrenamientos sobre dicho modelo, con el fin de evaluar su rendimiento en el contexto planteado en este estudio.

4.4.1. Recurso ClusterUY

El entrenamiento del modelo se realizó en el *Centro Nacional de Supercomputación* (**ClusterUY**) [43], una infraestructura compuesta por 45 nodos con sistema operativo *Linux CentOS 7*, interconectados mediante una red Ethernet de 10 Gbps.

Para este trabajo se solicitó un entorno de cómputo con 64 GB de memoria RAM y una **GPU NVIDIA Tesla P100**, equipada con 12 GB de memoria y 5384 núcleos CUDA. El entorno de ejecución se configuró mediante *Conda*, utilizando **CUDA 10.1** y **cuDNN 7**, junto con las bibliotecas indicadas en el entorno provisto por *Moliner et al.* [6].

Durante el entrenamiento se presentaron algunas limitaciones asociadas a la infraestructura del clúster, entre las que se destacan: la imposibilidad de asignar más de una GPU a un mismo trabajo, las restricciones de memoria de la GPU que condicionaron el tamaño del *batch* y la longitud de las secuencias de audio, y la necesidad de dividir los experimentos debido a que el sistema de colas no permite solicitar recursos por más de tres días consecutivos.

4.4.2. Entrenamientos

Una vez analizado en profundidad el modelo de dos etapas basado en *U-Net*, presentado en el Capítulo 3, se decidió realizar una serie de experimentos orientados a evaluar la capacidad del modelo bajo diferentes condiciones de ruido, adaptadas al contexto del presente trabajo: la restauración de grabaciones históricas de Lauro Ayestarán [3].

Con este objetivo, además del entrenamiento original implementado por Eloi Moliner en [6], se consideraron las siguientes variantes experimentales:

- **Modelo MusicNet + MagTapeDB:** entrenado utilizando la base *MusicNet* para las señales limpias y la base *MagTapeDB* como fuente de ruidos de cinta.
- **Modelo MusicNet + MagTapeDB + Gramófono:** entrenado con la base *MusicNet* para las señales limpias, y con las bases *MagTapeDB* y *Gramófono* combinadas como fuentes de ruido.

Es importante aclarar que la decisión de emplear ambas bases de ruidos responde al interés de analizar cómo la diversidad espectral y temporal de los distintos tipos de ruido analógico influye en la capacidad del modelo para generalizar y adaptarse a distintos escenarios de degradación sonora.

Por otra parte, no se modificó la base de datos de audios limpios, manteniéndose *MusicNet* como fuente principal. Esta elección se fundamenta en su accesibilidad

y extensión, además de la complejidad que implicaría construir una base alternativa de tamaño y diversidad comparables. Además, el uso en común de *MusicNet* también se justifica por la comparabilidad que ofrece con otros modelos entrenados sobre el mismo conjunto de grabaciones limpias. No obstante, esta decisión implica ciertas limitaciones, las cuales serán discutidas en la evaluación de los modelos entrenados, en el Capítulo 5.

La etapa de preprocesamiento de las bases de datos se mantuvo idéntica a la descrita en [6]. En el caso particular del conjunto conformado por *MagTapeDB* y *Gramófono*, ambos fueron preprocesados de manera independiente y, posteriormente, sus segmentos de ruido se combinaron de forma aleatoria. Una vez conformado el conjunto final, se realizó una división estratificada en un 70 % para entrenamiento y un 30 % para validación.

Como se mencionó anteriormente, debido a las limitaciones de recursos del *ClusterUY*, fue necesario restringir algunos de los hiperparámetros utilizados durante el entrenamiento. En particular, se redujo la duración de cada segmento de audio de 5 a 3 segundos, el tamaño del *batch* de 8 a 2, y la cantidad de épocas se estableció con un máximo de 150, aunque en cada corrida este valor depende exclusivamente del límite de tiempo asignado por el *ClusterUY*, cómo máximo tres días de uso. Es importante destacar que estas modificaciones tienen un impacto directo en el desempeño y la capacidad de generalización de los modelos resultantes del entrenamiento.

Durante el entrenamiento se monitorizaron dos métricas principales para evaluar el desempeño del modelo: la función de pérdida total y el *MAE* asociado a la segunda etapa de la U-Net (correspondiente a la sumatoria del segundo módulo en la Ecuación 3.3). Ambas métricas se calcularon tanto sobre el conjunto de entrenamiento como sobre el conjunto de validación. El seguimiento conjunto de estas curvas permitió analizar la evolución de la capacidad del modelo para aproximar los datos, así como identificar posibles indicios de sobreajuste o subajuste.

4.5. Evaluación de los modelos finales

En esta sección se presentan el procedimiento y los criterios empleados para evaluar el desempeño de los distintos modelos desarrollados en el contexto de la restauración de grabaciones musicológicas mediante técnicas de *denoising*. El objetivo principal es analizar comparativamente la capacidad de cada enfoque para atenuar el ruido sin degradar la calidad perceptual ni alterar la estructura armónica de las señales originales.

Para ello, se consideran tanto los métodos clásicos de sustracción espectral, en sus variantes tradicional y alternativa (*SS Clásico* y *SS Denoisify*), como los modelos basados en aprendizaje profundo entrenados con distintas combinaciones de bases de datos: *MusicNet*, *MagTapeDB* y *Gramófono*. De este modo, se busca evaluar el impacto que tiene la diversidad y naturaleza del conjunto de entrenamiento en la capacidad de generalización del modelo y en la preservación de las características propias del material sonoro restaurado. En resumen, los modelos sometidos a evaluación son los siguientes:

- **SS Clásico:** implementación tradicional del algoritmo de sustracción espectral.
- **SS Denoisify:** variante de la sustracción espectral basada en un esquema iterativo con modelado espectral y algoritmos de reducción de ruido musical.
- **DL Gramófono:** Modelo entrenado con la base de datos *MusicNet* y el conjunto de ruidos de *Gramófono*.
- **DL MagTapeDB:** Modelo entrenado con *MusicNet* y los ruidos provenientes de la base *MagTapeDB*.
- **DL MagTapeDB + Gramófono:** Modelo entrenado con *MusicNet*, *MagTapeDB* y *Gramófono*.

Para la evaluación se consideraron las grabaciones de la *Base de Música Personalizada* como señales limpias, y los audios de la base de datos *Grabaciones Analógicas de Cintas de Audio* como fuentes de ruido. Luego de una revisión detallada de esta última, se decidió excluir las grabaciones correspondientes al dispositivo *Revox C36*, debido a la presencia de un ruido tipo *buzz*, de carácter tonal y agudo, que resultaba perceptualmente dominante y poco representativo del tipo de ruido analógico que se busca estudiar en este trabajo.

Es importante destacar que para la evaluación final de los modelos no se consideró la base de datos *MusicNet*. Esto se debe a que uno de los objetivos centrales de este proyecto es analizar el desempeño de las distintas técnicas de *denoising* en un escenario más general y representativo del caso de estudio que motiva este trabajo: las grabaciones musicológicas de Lauro Ayestarán. Con este propósito, se construyó la *Base de Música Personalizada*, compuesta por señales altamente diversas entre sí, lo que permite evaluar los modelos en un contexto más amplio y exigente. Como algunas de estas señales presentan características similares a las incluidas en *MusicNet*, dicha base queda incorporada de forma implícita dentro de esta diversidad, evitando así sesgos hacia un conjunto específico y favoreciendo una evaluación más realista del rendimiento de cada modelo.

En primer lugar, cada una de las grabaciones de ruido de cinta, con una duración aproximada de 10 minutos, se segmentó en tramos de 30 segundos sin solapamiento, a fin de evitar posibles correlaciones entre las distintas combinaciones de señales limpias y ruidosas. Posteriormente, cada grabación limpia perteneciente a la *Base de Música Personalizada* se recortó considerando sus últimos 30 segundos, a los cuales se les asignó uno de los segmentos de ruido previamente definidos, conformando así las señales de evaluación contaminadas.

Este procedimiento se repitió para dos niveles de SNR, de 10 dB y 16 dB, siguiendo el criterio establecido en [4], donde dichos valores fueron seleccionados a partir de pruebas de escucha realizadas sobre grabaciones históricas de Lauro Ayestarán. La asignación de los segmentos de ruido se efectuó de manera que las grabaciones de los distintos dispositivos de cinta se distribuyeran uniformemente entre los grupos de audio de la base de datos, garantizando una representación equilibrada de las condiciones de evaluación.

Para la evaluación del rendimiento de los modelos, las señales limpias y ruidosas (en formato `WAV`) se organizaron en directorios independientes, clasificados por grupo y nivel de *SNR*, con el fin de aplicar posteriormente los distintos métodos de *denoising* y almacenar los resultados procesados en sus respectivas carpetas.

No obstante, este procedimiento presenta un posible inconveniente: al combinar una señal limpia con su correspondiente segmento de ruido, la suma puede exceder el rango normalizado de amplitud de las muestras, es decir, el intervalo $[-1, 1]$. Esto podría generar distorsión por saturación (*wrap around*) al guardar los archivos en formato `WAV`. Para prevenirlo, se realizó una normalización conjunta de ambas señales —limpia y ruidosa—, dividiendo cada una por el máximo absoluto entre ambas. Sea x la señal limpia y n el segmento de ruido asignado. La señal ruidosa se define como:

$$x_n = x + n. \tag{4.13}$$

Luego, se determinó un factor de normalización como el máximo absoluto entre ambas señales:

$$\text{norm\_factor} = \text{máx}\left(\text{máx}(|x|),\ \text{máx}(|x_n|)\right). \tag{4.14}$$

Finalmente, ambas se normalizaron mediante:

$$x \leftarrow \frac{x}{\text{norm\_factor}}, \qquad x_n \leftarrow \frac{x_n}{\text{norm\_factor}}. \tag{4.15}$$

De esta forma, se garantiza que la señal resultante permanezca dentro del rango permitido, evitando saturaciones sin alterar de manera significativa la SNR establecida. Además, se normalizó la señal limpia con el objetivo de mantener la coherencia de escalas, de modo que las versiones limpias y restauradas puedan compararse bajo las mismas condiciones de amplitud en cada una de las métricas utilizadas.

Cabe destacar que, si bien este reescalado introduce un error de cuantización mínimo, su impacto sobre la evaluación es despreciable frente a las diferencias significativas que se analizan entre los distintos modelos.

Las métricas seleccionadas para la evaluación final de los modelos fueron las variaciones $\Delta PAQM$ y $\Delta PEAQ$, definidas en las Ecuaciones 4.8 y 4.9, junto con el tiempo de ejecución correspondiente al proceso de restauración de cada modelo. Estas métricas permiten cuantificar tanto la mejora perceptual resultante como la eficiencia computacional de los distintos enfoques analizados.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 5

Análisis de resultados

El presente capítulo tiene por objetivo analizar y discutir los resultados obtenidos a partir de los distintos modelos y metodologías de reducción de ruido desarrollados durante este trabajo. A lo largo de los capítulos anteriores se abordaron las bases teóricas, el diseño de los modelos de sustracción espectral y aprendizaje profundo y los procedimientos experimentales empleados para su entrenamiento, validación y evaluación. En este capítulo se integran dichos elementos, presentando y analizando los resultados obtenidos, organizados en cuatro niveles de estudio complementarios que permiten una evaluación tanto cuantitativa como cualitativa del desempeño alcanzado por cada enfoque.

En primer lugar, se presentan los hiperparámetros finales obtenidos a partir de las búsquedas descritas en el Capítulo 4, los cuales definen las configuraciones óptimas encontradas para cada modelo de sustracción espectral. Estos valores fueron seleccionados en función de su desempeño en las métricas perceptuales y constituyen la base para las comparaciones realizadas en las siguientes secciones.

En segundo lugar, se examinan las curvas de aprendizaje, que reflejan la evolución de las métricas de entrenamiento y validación a lo largo de las épocas, permitiendo evaluar la convergencia y capacidad de generalización de los modelos de aprendizaje profundo, así como el impacto del conjunto de datos utilizado.

Luego, se realiza una evaluación objetiva mediante las métricas perceptuales, PEAQ y PAQM, que cuantifican la mejora en la calidad de las señales procesadas y permiten contrastar la efectividad de cada enfoque bajo diferentes condiciones de ruido y tipos de audio. Además, se incluye un estudio comparativo de los tiempos de procesamiento, con el fin de contextualizar los resultados obtenidos en términos de eficiencia computacional y viabilidad práctica de cada método.

Finalmente, se presenta una evaluación subjetiva, basada en la escucha crítica de las grabaciones restauradas, con el propósito de complementar el análisis numérico y obtener una apreciación perceptual del desempeño real de los modelos.

5.1. Búsqueda de hiperparámetros

A continuación se presentan los valores de los hiperparámetros obtenidos para los distintos módulos del algoritmo *SS Denoisify* y para *SS Clásico*. Asimismo, se realizó un análisis detallado del rendimiento alcanzado por cada una de las configuraciones óptimas.

5.1.1. Detector de inactividad

La configuración óptima obtenida para el detector de inactividad corresponde a la combinación de parámetros que alcanzó el mejor desempeño global según las métrica ponderada, especificada en la Subsección 4.3.4. Los valores medios y las desviaciones estándar del F_β-Score, *Precision*, *Recall* y RMSE se presentan en la Tabla 5.1. Además, los parámetros seleccionados para esta configuración se muestran en la Tabla 5.2.

Tabla 5.1: Resultados obtenidos para la configuración óptima del Detector de Inactividad. La metodología utilizada se detalla en Subsección 4.3.1.

Métrica	**Media**	**Desviación estándar**
F_β-Score (%)	68.65	17.28
Precision (%)	62.03	22.11
Recall (%)	91.84	9.18
RMSE	0.0047	0.0092

Tabla 5.2: Parámetros seleccionados para el Detector de Inactividad, entre los rangos de valores especificados en la Tabla 4.1.

Parámetro	**Valor**
`th_energy`	0.75
`th_zcr`	0.35
`th_he`	0.05
`zcr_hf_pct_cut`	0.90
`min_silence_len`	10
`min_sound_len`	25
`start_silence`	8
`end_silence`	1
`num_init_frames`	5

En primer lugar, los resultados obtenidos muestran que el detector de inactividad alcanzó un desempeño satisfactorio según el F_β-Score. Sin embargo, el análisis detallado de las métricas de *Precision* y *Recall* revela un comportamiento asimétrico: mientras que el *Recall* presenta un valor medio elevado, el *Precision* alcanza un valor medio sensiblemente menor.

Dado que el objetivo principal del módulo es maximizar el *Precision*, evitando clasificar como inactivos segmentos que en realidad contienen contenido sonoro, un valor relativamente bajo (62.03 %) evidencia un problema importante: el detector introduce una cantidad apreciable de falsos positivos, lo que implica que fragmentos con información musical pueden incorporarse erróneamente al cálculo del perfil

de ruido, afectando potencialmente la calidad de la restauración. Este comportamiento señala una limitación del módulo que deberá ser abordada y mejorada en posibles trabajos futuros.

Sin embargo, vale la pena destacar que el elevado *Recall* indica que la mayoría de los segmentos verdaderamente silenciosos sí son detectados correctamente, lo que beneficia la estimación de dicho perfil.

Luego, las desviaciones estándar observadas en la Tabla 5.1 indican que el desempeño del detector varía de forma considerable entre grabaciones. Esto sugiere que la efectividad del módulo depende fuertemente de las características particulares de cada señal —como su dinámica, instrumentación o presencia de transitorios—, lo que repercute en la estabilidad de la detección.

Cabe destacar que los resultados obtenidos para la métrica *RMSE* —una media de **0.0047** y una desviación estándar de **0.0092**— son satisfactorios para el objetivo planteado. Estos valores reflejan una estimación consistente del perfil de ruido, incluso considerando que el desempeño puede verse afectado por errores en la *Precision*, ya que la inclusión incorrecta de *frames* activos dentro del perfil de ruido introduce cierta distorsión en la estimación.

Por otro lado, dentro de los parámetros seleccionados resulta especialmente interesante analizar los valores asociados a las longitudes mínimas de *frames* requeridas para considerar un segmento como activo o inactivo. A una frecuencia de muestreo de 44,1 kHz, la duración mínima establecida para un segmento de actividad es de aproximadamente 330 ms, mientras que la correspondiente a un segmento inactivo es cercana a 150 ms, es decir, casi la mitad de la anterior.

Asimismo, los márgenes temporales definidos al inicio y al final de cada segmento de silencio muestran una asimetría significativa: los *frames* de margen inicial corresponden aproximadamente a 130 ms, mientras que los del margen final representan cerca de 50 ms. Esta diferencia sugiere que, en términos prácticos, un segmento activo tiende a presentar una transición más gradual al finalizar que al iniciarse, lo cual indica que la aparición de actividad en la señal suele ser más abrupta que su finalización.

Para ilustrar con mayor precisión el desempeño alcanzado, se seleccionó una señal ruidosa del conjunto de evaluación que obtuvo un resultado particularmente desfavorable. Sobre esta señal se analizaron en detalle sus características tanto temporales como espectrales. En la Figura 5.1 se representan la STE, la ZCR y la MHF. Por otro lado, en la Figura 5.2 se muestran los resultados del proceso de detección junto con los perfiles de ruido estimados y etiquetados.

La Figura 5.1 debe interpretarse considerando que la señal analizada corresponde a los últimos 30 segundos de una grabación musical de la *Base de Música Personalizada*. Por lo tanto, es coherente observar que, a partir de aproximadamente los 20–22 segundos, tanto la STE como la MHF disminuyen de manera significativa, reflejando el final natural de la pieza musical y la desaparición progresiva de sus componentes estructurales.

Es importante destacar que los umbrales utilizados para la STE y la MHF son extremadamente bajos. Esto se debe a que, incluso para una SNR = 10 dB, la energía y la magnitud espectral del ruido de cinta son considerablemente menores

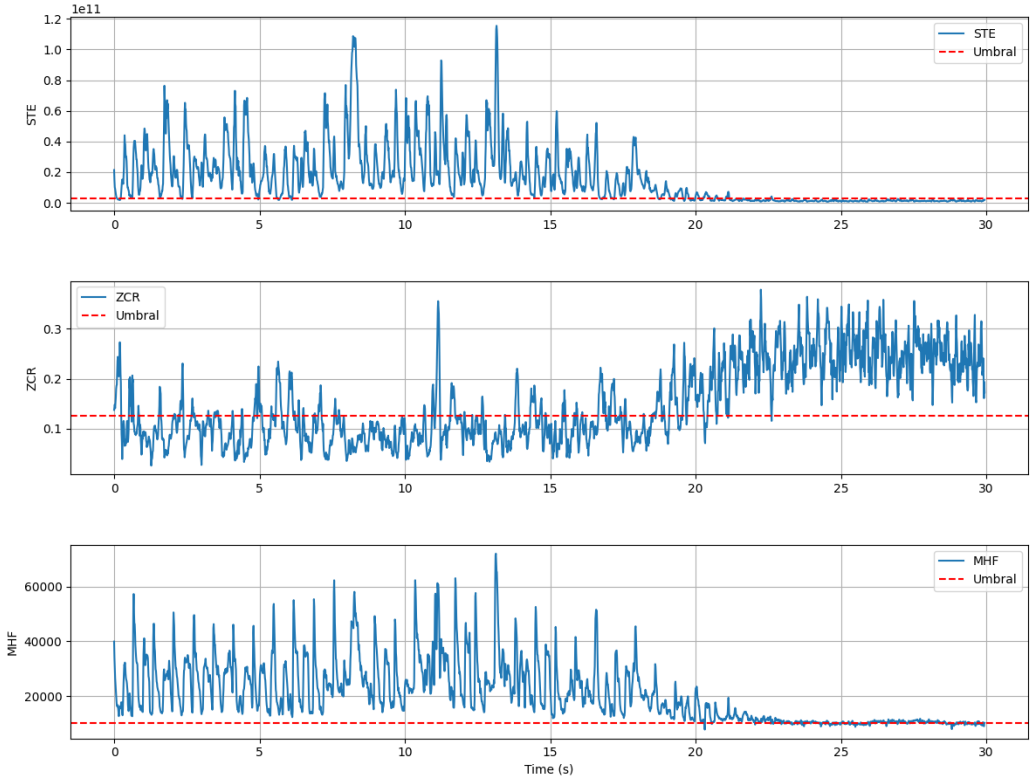

Figura 5.1: Evolución temporal de las tres métricas utilizadas en el detector de inactividad: STE, ZCR y MHF, junto con sus umbrales respectivos. La figura ilustra cómo, hacia los últimos segundos del fragmento ($\sim$ 20–22 s), las métricas basadas en energía y magnitud espectral descienden de manera significativa debido al final natural de la pieza musical, mientras que la ZCR aumenta en ausencia de contenido tonal, reflejando la presencia dominante del ruido de cinta en altas frecuencias.

que las de la señal musical. En contraste, el umbral de la ZCR es más elevado, lo cual resulta esperable dado que el ruido analógico presenta variaciones de signo mucho más frecuentes que la señal original, aun cuando su potencia es baja.

Un comportamiento particularmente ilustrativo se observa hacia el final del fragmento, cuando la música se extingue: la ZCR aumenta de manera sostenida. Esto indica que, en ausencia del contenido armónico de la señal original, prevalece únicamente el ruido de cinta. Este incremento en la ZCR coincide con la hipótesis discutida en el Apéndice A, donde se plantea que el ruido de cinta considerado en este trabajo posee una componente espectral que se intensifica en las bandas altas, lo que naturalmente incrementa su tasa de cruces por cero.

Por otra parte, la señal elegida obtuvo valores de *Precision*, *Recall* y *RMSE* iguales a **34.27**, **68.69** y **0.0013**, respectivamente. Como se mencionó anteriormente, estos resultados no son satisfactorios y esto se puede ver reflejado directamente en la Figura 5.2. En el primer panel se observa que una proporción considerable de los segmentos activos fue clasificada erróneamente como inactiva, lo cual explica el valor relativamente bajo de *Precision*. Al mismo tiempo, puede apreciarse

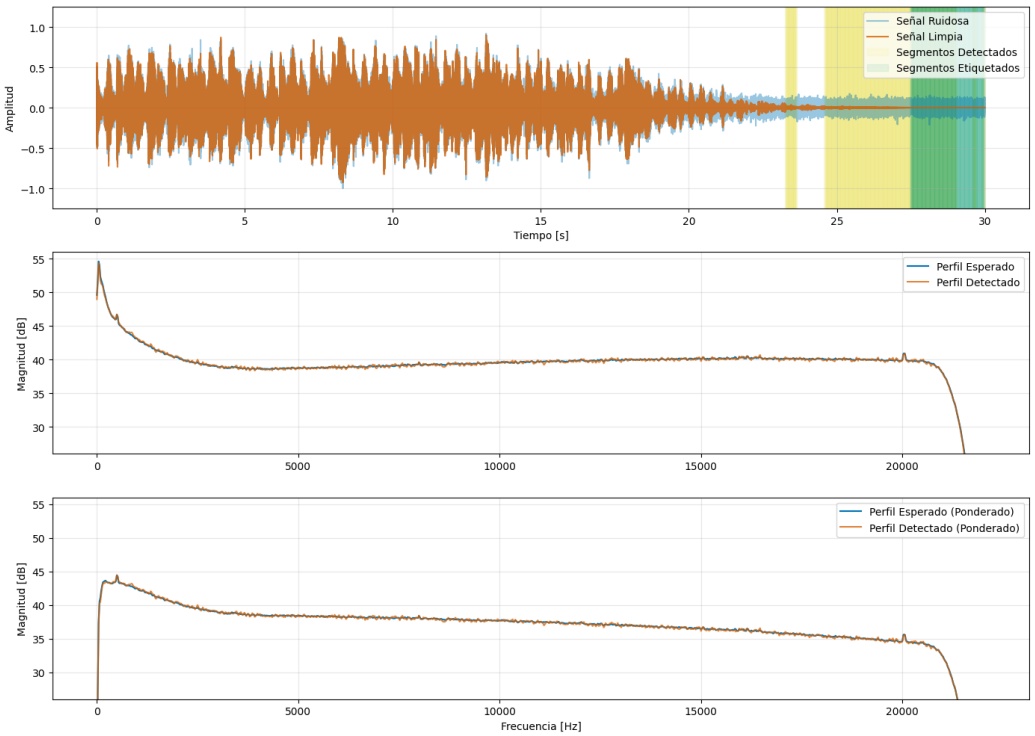

Figura 5.2: Desempeño del detector de inactividad para una señal del conjunto de evaluación. En el panel superior se muestran la señal ruidosa, la señal limpia y los segmentos inactivos detectados en comparación con los segmentos etiquetados manualmente. En los paneles centrales e inferiores se ilustran los perfiles espectrales de ruido esperados y detectados, tanto sin ponderación como aplicando A-Weighting. La figura permite visualizar simultáneamente los aciertos y fallos en la detección temporal, así como la elevada precisión alcanzada en la estimación espectral del ruido.

que la mayoría de los segmentos etiquetados manualmente como inactivos fueron detectados correctamente por el algoritmo, lo que se corresponde con un *Recall* significativamente más alto.

El bajo desempeño del *Precision* puede justificarse a partir del comportamiento temporal de la señal: cuando la canción comienza a finalizar (aproximadamente a partir de los 22 s), su potencia se vuelve muy baja en comparación con la del ruido de cinta presente a un SNR = 10 dB. En estas condiciones, el algoritmo no logra distinguir adecuadamente la señal limpia en esos tramos, interpretando dichos segmentos como inactivos.

Este comportamiento también puede observarse en la Figura 5.1, donde alrededor de los 22 s se evidencia un valor de ZCR excesivamente elevado, característico del ruido, en lugar de un ZCR más reducido que correspondería a la señal original. Debido a que dicho valor queda por encima del umbral utilizado para identificar tramos inactivos, estos *frames* son erróneamente clasificados pese a contener actividad relevante de la señal.

Por otro lado, los dos paneles inferiores de Figura 5.2 muestran los perfiles es-

pectrales de ruido —esperado y detectado— tanto sin ponderación como aplicando A-Weighting. En ambos casos se observa una coincidencia notable entre los perfiles, lo cual respalda el valor extremadamente bajo del *RMSE* obtenido y confirma que, pese a los errores en la detección temporal, la estimación espectral del ruido sigue siendo altamente precisa.

5.1.2. Sustracción espectral

El rendimiento obtenido por la configuración óptima del algoritmo *SS Clásico* se resume en la Tabla 5.3, donde se presentan los valores medios y las desviaciones estándar de ΔPAQM, ΔPEAQ y la SNR estimada. Los hiperparámetros que permitieron obtener estos resultados se detallan en la Tabla 5.5.

En el caso del algoritmo *SS Denoisify*, la configuración óptima alcanzada muestra un desempeño consistente según las mismas métricas, cuyos valores se reportan en la Tabla 5.4. Los hiperparámetros asociados a esta configuración se presentan en la Tabla 5.6.

Tabla 5.3: Rendimiento obtenido para la configuración óptima del algoritmo *SS Clásico*.

Métrica	**Media**	**Desviación estándar**
ΔPAQM	1.94	0.91
ΔPEAQ	0.50	0.73
SNR (dB)	22.32	3.10

Tabla 5.4: Rendimiento obtenido para la configuración óptima del algoritmo *SS Denoisify*.

Métrica	**Media**	**Desviación estándar**
ΔPAQM	1.87	0.93
ΔPEAQ	0.33	0.45
SNR (dB)	22.35	3.24

Tabla 5.5: Configuración óptima encontrada para el algoritmo *SS Clásico*, a partir de los rangos de valores elegidos en la Tabla 4.2.

Hiperparámetro	**Valor**
`alpha`	0.99
`beta`	0.01
`rho`	0.15

Tanto la versión clásica como *SS Denoisify* lograron mejoras perceptuales positivas, reflejadas en los valores de $\Delta PAQM$ y $\Delta PEAQ$. No obstante, la sustracción espectral clásica presentó un desempeño superior, alcanzando incrementos promedio mayores en ambas métricas en comparación con *SS Denoisify*. Dado que la SNR estimada resultó prácticamente equivalente en ambos casos, las diferencias

Tabla 5.6: Configuración óptima encontrada para el algoritmo *SS Denoisify*, a partir de los rangos de valores elegidos en la Tabla 4.3.

Hiperparámetro	Valor
n_iter	30
alpha	0.10
beta	0.93
rho	0.01
sm_keep_pct	0.50
peak_thresh	-60

observadas no se deben a la cantidad de ruido atenuado, sino al impacto que cada algoritmo introduce sobre la estructura temporal y espectral de la señal.

Las desviaciones estándar relativamente elevadas en todas las métricas indican que el rendimiento de ambos enfoques depende fuertemente de las características específicas de cada señal musical, un comportamiento esperable dada la diversidad armónica y dinámica presente en las grabaciones evaluadas.

Un aspecto particularmente relevante de estos resultados son las configuraciones óptimas encontradas para cada algoritmo. En el caso de la sustracción espectral clásica, los parámetros seleccionados corresponden sistemáticamente a los valores más extremos dentro de los rangos evaluados: un $\alpha = 0{,}99$ que maximiza la cantidad de ruido sustraído, un $\beta = 0{,}01$ que fija un piso espectral extremadamente bajo, y un $\rho = 0{,}15$ que aplica un filtro pasabajos más agresivo. Esta combinación sugiere que, para el tipo de ruido analógico considerado, la versión clásica del algoritmo se beneficia de una estrategia extremadamente agresiva en la eliminación del ruido, incluso a costa de un mayor riesgo de distorsión. Además, el hecho de que los valores óptimos se encuentren en los extremos del espacio de búsqueda sugiere que la configuración verdaderamente óptima podría ubicarse más allá de los límites evaluados, lo cual puede ser interesante para explorar en trabajos futuros.

En contraste, la configuración óptima de *SS Denoisify* presenta un comportamiento significativamente distinto: prioriza el uso del número máximo de iteraciones (`n_iter` = 30), una proporción moderada de iteraciones dedicadas al modelado espectral (`sm_keep_pct` = 0.50), y un umbral de detección de picos sinusoidales muy bajo (`peak_thresh` = -60 dB), lo que implica una detección amplia de componentes tonales durante la síntesis del modelado sinusoidal. Asimismo, los valores óptimos de $\alpha = 0{,}10$ y $\beta = 0{,}93$ indican un enfoque considerablemente menos agresivo tanto en la sustracción como en el piso espectral. En conjunto, estos parámetros sugieren que *SS Denoisify* obtiene su mejor rendimiento cuando atenúa el ruido de forma más moderada y delega un papel preponderante al modelado sinusoidal y a la sustracción iterativa.

Es importante mencionar que el análisis detallado del rendimiento de ambos algoritmos se desarrollará en profundidad en las secciones siguientes.

5.1.3. Algoritmo de reducción de ruido musical

Los resultados obtenidos para la configuración óptima del algoritmo se resumen en la Tabla 5.7, donde se presentan los valores medios y las desviaciones estándar de ΔPEAQ y ΔPAQM. La configuración de hiperparámetros asociada a estos resultados se muestra en la Tabla 5.8.

Tabla 5.7: Rendimiento obtenido para la configuración óptima del algoritmo.

Métrica	**Media**	**Desviación estándar**
ΔPEAQ	-0.0750	1.1120
ΔPAQM	0.0109	0.1790

Tabla 5.8: Configuración óptima encontrada para el algoritmo de reducción de ruido musical, a partir de los rangos de valores elegidos en la Tabla 4.4.

Hiperparámetro	**Valor**
`mn_nfft`	256
`mn_thresh_db`	-25
`mn_win_len`	44

En ambas tablas se puede destacar que el desempeño alcanzado por la configuración óptima del algoritmo de reducción de ruido musical no es favorable. Los valores obtenidos para $\Delta PEAQ$ y $\Delta PAQM$ son prácticamente nulos en promedio, lo que indica que el módulo no aporta mejoras significativas al proceso de *denoising*. Además, al igual que en los casos anteriores, las desviaciones estándar asociadas a ambas métricas son considerablemente elevadas, lo que refleja una fuerte dependencia del rendimiento respecto de las características particulares de cada grabación. En conjunto, estos resultados sugieren que, bajo las configuraciones evaluadas, el módulo de reducción de ruido musical no logra contribuir de forma consistente a la restauración de las señales. Debido a esta razón, en las evaluaciones y análisis posteriores no se consideró el último módulo de la implementación.

5.2. Curvas de aprendizaje

En los capítulos anteriores se presentaron los fundamentos teóricos y metodológicos que sustentan el desarrollo y entrenamiento de los modelos de denoising basado en redes neuronales profundas.

En particular, se entrenaron dos versiones del modelo: una utilizando exclusivamente la base de datos de ruidos de cinta MagTapeDB, y otra combinando dicha base con la colección de ruidos de gramófono empleada por Moliner y Välimäki [6]. El objetivo de esta comparación fue evaluar si un modelo entrenado con una base específica para el tipo de ruido presente en las grabaciones de cinta logra un mejor desempeño que uno entrenado con una combinación más diversa de ruidos, analizando así el efecto de la generalización frente a la especialización del conjunto de entrenamiento.

En este contexto, las curvas de aprendizaje presentadas a continuación permiten analizar la evolución del error durante el proceso de entrenamiento, tanto en el conjunto de entrenamiento como en el de validación. Su observación resulta fundamental para evaluar la convergencia del modelo, su capacidad de generalización y el impacto de la base de datos empleada en el desempeño final.

Las Figuras 5.3 a 5.6 muestran la evolución de la pérdida y del error absoluto medio registrados durante el proceso de entrenamiento para ambos modelos considerados. En cada caso se presentan las métricas de entrenamiento y validación a lo largo de las épocas.

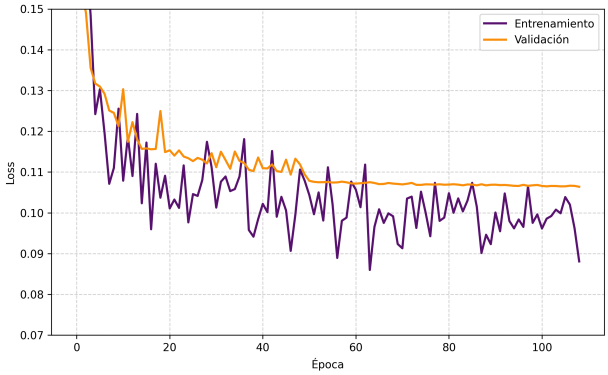

Figura 5.3: Evolución de la pérdida durante el entrenamiento y la validación para el modelo entrenado con la base MagTapeDB.

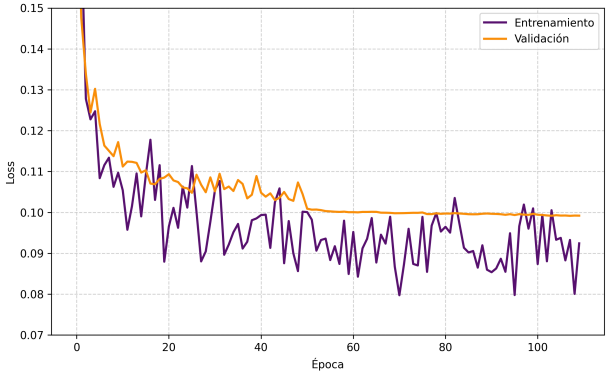

Figura 5.4: Evolución de la pérdida durante el entrenamiento y la validación para el modelo entrenado con la base combinada MagTapeDB + Gramófono.

Al comparar las curvas correspondientes a cada modelo, se observa que la evolución de la pérdida y del MAE presentan comportamientos muy similares dentro de una misma base de datos. En ambos casos, las curvas siguen una tendencia decreciente durante las primeras épocas y una posterior estabilización en validación, lo que indica que ambas métricas reflejan de manera coherente la dinámica del proceso de aprendizaje, aunque cuantifican magnitudes distintas.

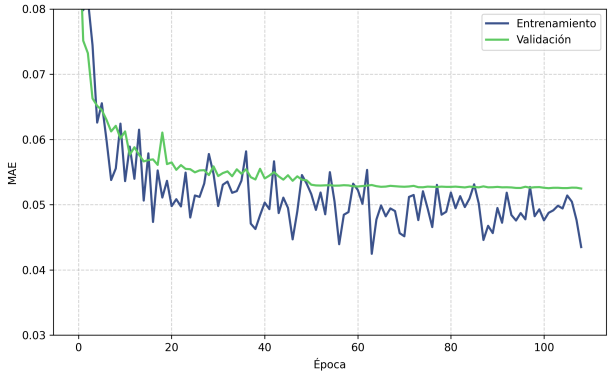

Figura 5.5: Evolución del MAE durante el entrenamiento y la validación para el modelo entrenado con la base MagTapeDB.

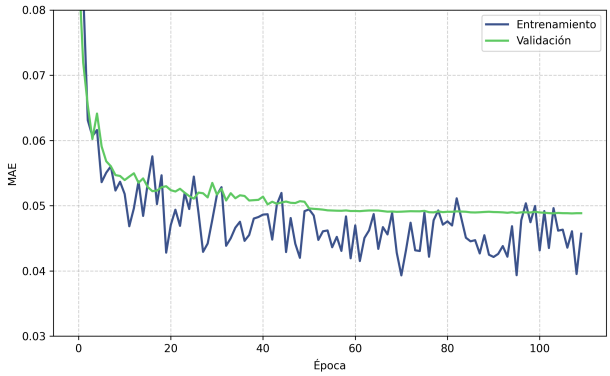

Figura 5.6: Evolución del MAE durante el entrenamiento y la validación para el modelo entrenado con la base combinada MagTapeDB + Gramófono.

Tal como se detalla en la Subsección 3.2.2, la función de pérdida utilizada durante el entrenamiento está definida como la suma de dos términos: el MAE de la salida intermedia $\hat{Y}_1$ y el MAE de la salida final $\hat{Y}_2$ del modelo. En cambio, la métrica *MAE* reportada por `TensorBoard` corresponde únicamente al error absoluto medio de la segunda salida $\hat{Y}_2$.

Debido a esta relación directa ambas curvas siguen una tendencia paralela, diferenciándose más en su escala numérica que en su forma.

A partir de la época 50, tanto las curvas de pérdida como las de error absoluto medio muestran una clara estabilización en el conjunto de validación. Este comportamiento coincide con la reducción automática de la tasa de aprendizaje implementada en el código, donde el *learning rate* se reduce en un factor de 10 cada 50 épocas. Al disminuir el tamaño de los pasos del optimizador en cada actualización, el modelo realiza un ajuste más fino, lo que se traduce en una convergencia más lenta pero más estable de las métricas de validación.

La mayor estabilidad observada en las curvas de validación, en comparación con las de entrenamiento, podría estar asociada a la forma en que se construyen

ambos conjuntos de datos. Durante el entrenamiento, los segmentos se contaminan en tiempo real mediante la adición de ruido, asignándoles en cada iteración una SNR aleatoria dentro del intervalo 2–20 dB. Esta variabilidad hace que, dentro de un mismo lote, coexistan ejemplos muy ruidosos y otros con ruido moderado, lo que podría introducir una dispersión importante en los valores de la pérdida y dar lugar a curvas de entrenamiento más inestables. En cambio, el conjunto de validación se genera una única vez antes de iniciar el entrenamiento: a cada segmento se le asigna una SNR aleatoria, también dentro del rango 2–20 dB, y esa configuración se mantiene fija a lo largo de todas las épocas. De esta forma, la evaluación se realiza siempre sobre los mismos ejemplos y niveles de ruido, lo que tendería a producir curvas de validación más suaves y estables.

Además, el rango de SNR considerado (2–20 dB) es relativamente amplio y abarca condiciones muy distintas de degradación. Para valores próximos a 2 dB, el error esperado suele ser significativamente mayor que para SNR altos, de modo que la combinación de todos estos casos en un único proceso de entrenamiento podría contribuir a la dispersión observada en las curvas.

Por otro lado, al contrastar los resultados entre modelos, se observa que el modelo entrenado únicamente con la base MagTapeDB presenta valores de pérdida de validación ligeramente superiores a los del modelo entrenado con la base combinada MagTapeDB + Gramófono. Sin embargo, esta diferencia no resulta concluyente dado que puede deberse a variaciones inherentes al proceso de entrenamiento, como el ordenamiento aleatorio de los datos o la distribución de los ejemplos en cada conjunto.

Finalmente, en el presente trabajo las condiciones de entrenamiento difieren sustancialmente respecto a las del estudio original de Moliner y Välimäki. No se dispone de las curvas de entrenamiento y validación correspondientes al modelo original, entrenado exclusivamente sobre la base de datos de gramófono, sin embargo, es razonable suponer que los autores contaban con un *hardware* más potente, dado que durante la replicación del entrenamiento en este trabajo se registraron errores de asignación de memoria al intentar utilizar configuraciones equivalentes. Como se mencionó en el Capítulo 4, el entrenamiento se realizó en una GPU con 12 GB de VRAM, lo que obligó a reducir el tamaño de lote y ajustar otros hiperparámetros para adaptarse a la capacidad de memoria. Estas limitaciones, detalladas en 4.4.2 , pueden explicar en parte la inestabilidad observada en las curvas de entrenamiento.

5.3. Análisis objetivo de los modelos

Con el objetivo de evaluar cuantitativamente el desempeño de los distintos métodos de reducción de ruido, se procesaron todos los casos de prueba y se calcularon las variaciones promedio de las métricas objetivas de calidad perceptual PEAQ y PAQM, descritas previamente en las Secciones 4.2.3 y 4.2.4.

Previo al comienzo del análisis, es importante destacar que la elección de estas dos métricas no fue arbitraria: ambas ofrecen una estimación objetiva del impacto

perceptual que introduce un proceso, pero desde perspectivas ligeramente diferentes y, por tanto, complementarias.

Durante el análisis experimental se observó que PEAQ tiende a favorecer enfoques de denoising más agresivos, en los cuales la supresión del ruido es prioritaria, incluso a costa de introducir cierta distorsión residual en la señal limpia. Por el contrario, PAQM penaliza con mayor severidad ese tipo de distorsiones, valorando en cambio una preservación más fiel del contenido armónico y tímbrico original, aun cuando el ruido residual es algo mayor.

Esta diferencia de comportamiento hace que la combinación de ambas métricas proporcione una visión más equilibrada y representativa del rendimiento real de los modelos.

5.3.1. Desempeño general

La Figura 5.7 resume el comportamiento global de cada modelo, promediando todas las condiciones de SNR y tipos de audio. La Tabla 2.1, acompaña a esta figura y presenta los valores promedio y desviaciones estándar de las métricas ΔPEAQ y ΔPAQM para cada método, lo cual aporta una visión cuantitativa del desempeño general.

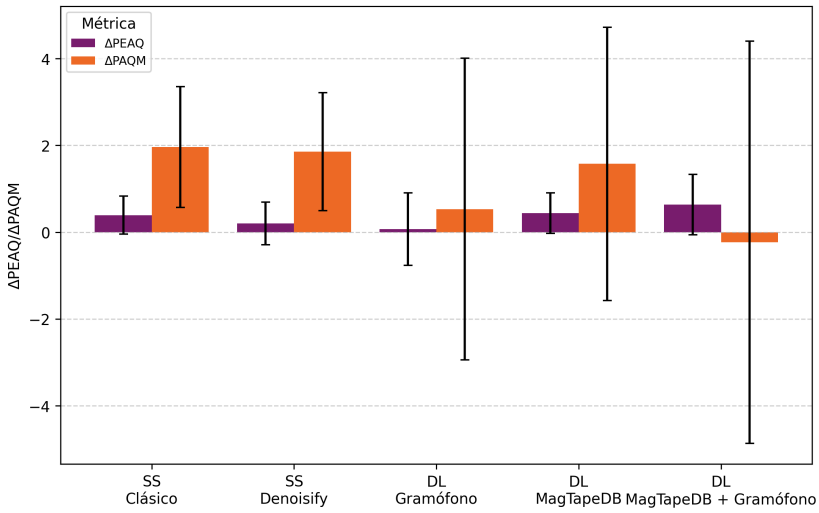

Figura 5.7: Resumen global del desempeño de los distintos métodos de reducción de ruido, evaluados mediante las métricas objetivas ΔPEAQ y ΔPAQM. La figura ilustra el contraste entre las técnicas clásicas de sustracción espectral —que muestran resultados consistentes y relativamente estables— y los modelos basados en aprendizaje profundo, cuyo rendimiento evidencia una mayor variabilidad y una fuerte dependencia del conjunto de entrenamiento.

En primer lugar, los métodos de sustracción espectral exhiben desempeños muy similares entre sí, con valores promedio comparables en ambas métricas y variabilidades moderadas. Esto sugiere que, pese a las diferencias metodológicas entre ambas variantes, su comportamiento general frente a las señales evaluadas es consistente.

Tabla 5.9: Resumen global de los valores promedio y desviación estándar de las métricas objetivas por modelo.

Método	ΔPEAQ	$\sigma_{\Delta\text{PEAQ}}$	ΔPAQM	$\sigma_{\Delta\text{PAQM}}$
SS Clásico	0.392	**0.439**	**1.963**	1.390
SS Denoisify	0.202	0.489	1.856	**1.360**
DL Gramófono	0.070	0.835	0.532	3.477
DL MagTapeDB	0.439	0.465	1.575	3.151
DL MagTapeDB + Gramófono	**0.638**	0.696	-0.230	4.635

En contraste, los modelos basados en aprendizaje profundo muestran una divergencia clara entre sí, lo que indica una fuerte dependencia respecto del conjunto de entrenamiento utilizado. En particular, el modelo *DL MagTapeDB* tiende a obtener mejores resultados en ΔPEAQ y ΔPAQM que los modelos entrenados con *Gramófono*, lo cual es coherente con las condiciones de la evaluación: todas las pruebas se realizaron exclusivamente con ruido de cinta proveniente de la base *Analog Audio Tape Recordings*. Dado que las características espectrales y temporales del ruido de cinta difieren del ruido propio de grabaciones en gramófono, es razonable que un modelo entrenado con este último no logre generalizar adecuadamente al dominio del ruido de cinta. En cambio, el modelo entrenado con *MagTapeDB* dispone de ejemplos representativos del tipo de ruido presente en la evaluación, lo que explica su desempeño superior.

Aun así, en términos generales, los métodos de aprendizaje profundo presentan un rendimiento significativamente inferior en comparación con las técnicas de sustracción espectral, especialmente en la métrica ΔPAQM, donde muestran medias sensiblemente menores —incluso negativas en el caso del modelo *DL MagTapeDB+Gramófono*— y desviaciones estándar mucho más elevadas. Esto indica no solo una menor eficacia en la restauración de la señal, sino también una mayor inestabilidad frente a la variabilidad del conjunto de evaluación. Por su parte, en ΔPEAQ los modelos de aprendizaje profundo pueden alcanzar valores competitivos; el modelo *DL MagTapeDB+Gramófono* es el que obtiene el mejor desempeño global en esta métrica. Sin embargo, esta mejora viene acompañada nuevamente de una variabilidad considerable, lo que limita su fiabilidad en escenarios más generales.

Resulta relevante señalar que, en general, el margen de mejora en la restauración de las señales es mayor que el margen de deterioro: las variaciones positivas en las métricas aparecen con mayor frecuencia que las negativas. No obstante, los modelos basados en aprendizaje profundo exhiben niveles de degradación mucho más pronunciados que las técnicas de sustracción espectral, evidenciando una mayor vulnerabilidad frente a señales que se desvían del dominio visto durante el entrenamiento. Esto refuerza la importancia tanto del conjunto de entrenamiento como de la capacidad de generalización del modelo, especialmente cuando se trabajan señales con características acústicas diferentes.

5.3.2. Desempeño por SNR

La Figura 5.8 y la Figura 5.9 ilustran la evolución de las métricas ΔPEAQ y ΔPAQM para cada modelo bajo condiciones de 10 y 16 dB de SNR. Complementariamente, las Tablas 5.10 y 5.11 presentan los valores promedio y las desviaciones estándar correspondientes a cada caso.

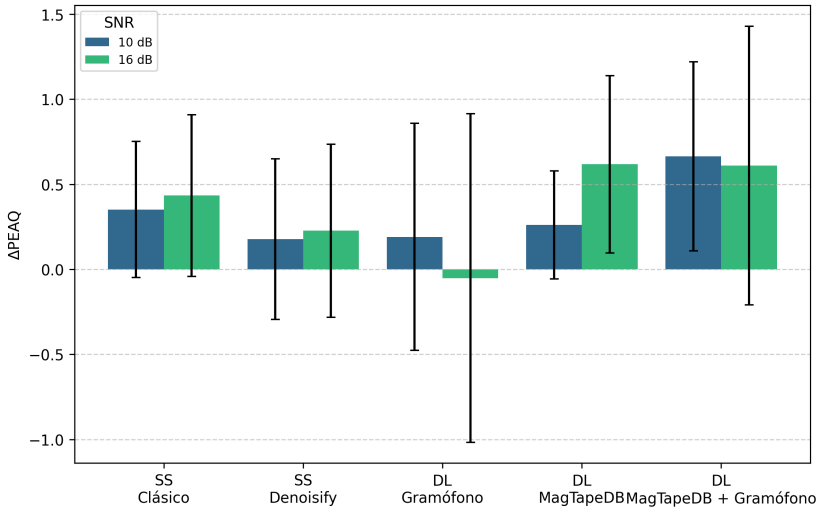

Figura 5.8: Variación promedio de ΔPEAQ para cada modelo bajo las dos condiciones de SNR consideradas (10 y 16 dB). La figura permite observar cómo se modifica la calidad perceptual estimada según el nivel de ruido de entrada y comparar la sensibilidad de cada método frente a esta condición.

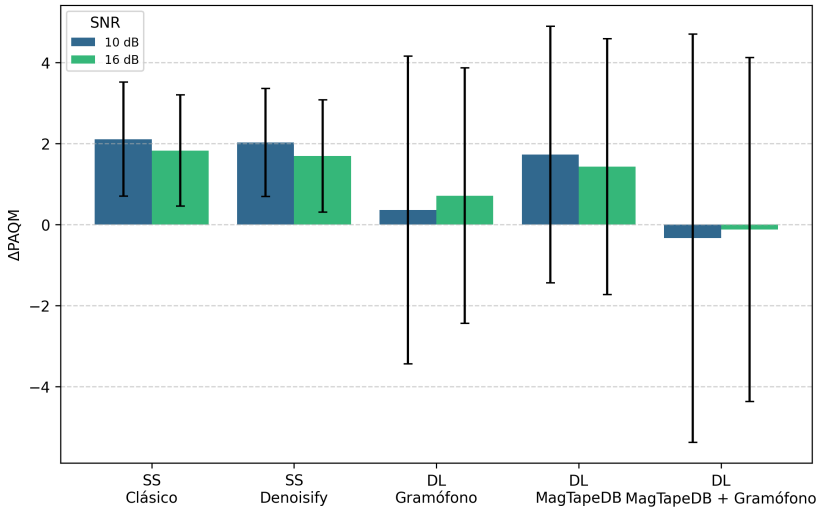

Figura 5.9: Variación promedio de ΔPAQM por modelo para SNR de 10 y 16 dB. Se muestra cómo cada técnica preserva o degrada la calidad perceptual según el nivel de ruido de la señal ruidosa, permitiendo identificar patrones de estabilidad o sensibilidad frente al SNR.

Tabla 5.10: Resultados promedio y desviación estándar de las métricas objetivas para un SNR de 10 dB.

Método	ΔPEAQ	$\sigma_{\Delta\text{PEAQ}}$	ΔPAQM	$\sigma_{\Delta\text{PAQM}}$
SS Clásico	0.351	0.400	**2.104**	1.409
SS Denoisify	0.178	0.473	2.023	**1.331**
DL Gramófono	0.191	0.668	0.356	3.800
DL MagTapeDB	0.261	**0.317**	1.725	3.169
DL MagTapeDB + Gramófono	**0.665**	0.555	-0.338	5.038

Tabla 5.11: Resultados promedio y desviación estándar de las métricas objetivas para un SNR de 16 dB.

Método	ΔPEAQ	$\sigma_{\Delta\text{PEAQ}}$	ΔPAQM	$\sigma_{\Delta\text{PAQM}}$
SS Clásico	0.434	**0.476**	**1.823**	**1.371**
SS Denoisify	0.227	0.509	1.689	1.383
DL Gramófono	-0.051	0.965	0.708	3.153
DL MagTapeDB	**0.618**	0.521	1.425	3.159
DL MagTapeDB + Gramófono	0.611	0.819	-0.122	4.244

Tanto *SS Clásico* como *SS Denoisify* presentan un comportamiento estable entre las dos condiciones analizadas. Las variaciones entre los valores medios de ΔPEAQ y ΔPAQM para ambos casos son relativamente pequeñas, y sus desviaciones estándar se mantienen acotadas, lo que indica que la eficacia de estas técnicas no depende principalmente del SNR del audio de entrada. Esto es coherente con los algoritmos de sustracción espectral, cuyo funcionamiento no incorpora explícitamente el valor del SNR para el procesamiento de la señal ruidosa.

Por el contrario, los modelos basados en aprendizaje profundo denotan un comportamiento mucho más irregular y sin una tendencia clara asociada al SNR. Tanto *DL Gramófono* como *DL MagTapeDB* y su combinación muestran variaciones significativas entre 10 y 16 dB, tanto en el sentido de mejora o deterioro como en la magnitud de la dispersión. Esta inestabilidad puede explicarse por la fuerte dependencia de estos modelos respecto de los datos utilizados durante el entrenamiento. En particular, las redes, como se describe en la Subsección 3.2.1, se entrenaron con un rango considerablemente amplio de SNR —entre 2 y 20 dB—, por lo que el modelo debe aprender simultáneamente a manejar niveles de ruido muy distintos, lo cual dificulta una generalización adecuada para condiciones específicas como las evaluadas en este trabajo.

Estos resultados muestran que las técnicas clásicas mantienen un desempeño más predecible y robusto frente a cambios en el SNR, mientras que los modelos neuronales son más sensibles al rango de condiciones visto en entrenamiento y, por ello, muestran una estabilidad mucho menor.

5.3.3. Desempeño por categoría de audio

Las Figuras 5.10 y 5.11, junto con las Tablas 5.13–5.15, resumen el comportamiento de cada modelo según la categoría de contenido musical presente en la señal. Este análisis permite evaluar hasta qué punto la característica espectral y temporal de las grabaciones de audio influye en la eficacia del algoritmo de *denoising*.

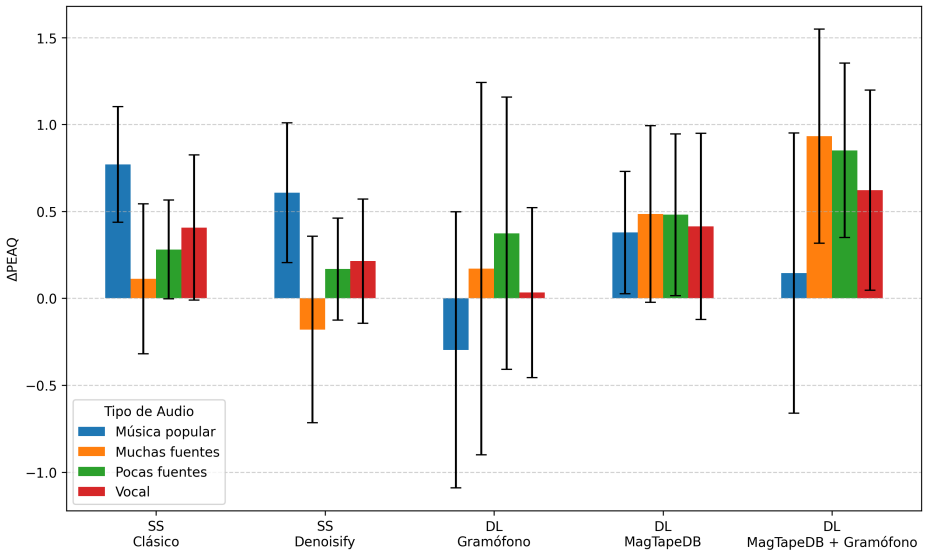

Figura 5.10: Comparación de ΔPEAQ por tipo de contenido musical. Las barras indican valores promedio y las líneas de error su desviación estándar.

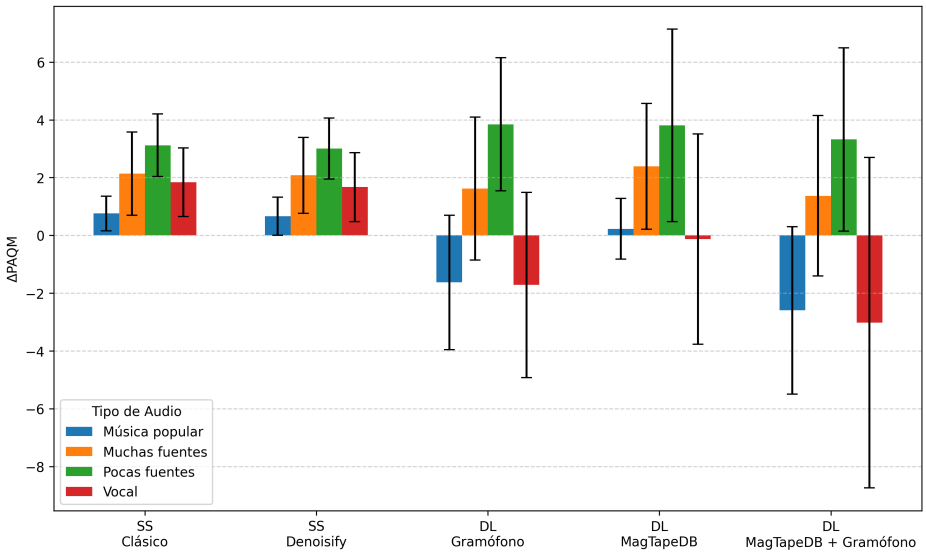

Figura 5.11: Comparación de ΔPAQM por tipo de contenido musical. Las barras indican valores promedio y las líneas de error su desviación estándar.

Tabla 5.12: Resultados promedio y desviación estándar de las métricas objetivas para el tipo de audio Música popular.

Método	ΔPEAQ	$\sigma_{\Delta\text{PEAQ}}$	ΔPAQM	$\sigma_{\Delta\text{PAQM}}$
SS Clásico	**0.770**	**0.332**	**0.758**	**0.596**
SS Denoisify	0.607	0.402	0.663	0.659
DL Gramófono	-0.297	0.794	-1.625	2.326
DL MagTapeDB	0.379	0.352	0.226	1.050
DL MagTapeDB + Gramófono	0.145	0.806	-2.595	2.899

Tabla 5.13: Resultados promedio y desviación estándar de las métricas objetivas para el tipo de audio Muchas fuentes.

Método	ΔPEAQ	$\sigma_{\Delta\text{PEAQ}}$	ΔPAQM	$\sigma_{\Delta\text{PAQM}}$
SS Clásico	0.112	**0.432**	2.138	1.444
SS Denoisify	-0.180	0.537	2.081	**1.314**
DL Gramófono	0.170	1.071	1.625	2.475
DL MagTapeDB	0.484	0.508	**2.393**	2.179
DL MagTapeDB + Gramófono	**0.932**	0.616	1.372	2.779

Tabla 5.14: Resultados promedio y desviación estándar de las métricas objetivas para el tipo de audio Pocas fuentes.

Método	ΔPEAQ	$\sigma_{\Delta\text{PEAQ}}$	ΔPAQM	$\sigma_{\Delta\text{PAQM}}$
SS Clásico	0.281	**0.284**	3.117	1.084
SS Denoisify	0.168	0.293	3.007	**1.052**
DL Gramófono	0.374	0.783	**3.845**	2.303
DL MagTapeDB	0.481	0.465	3.807	3.331
DL MagTapeDB + Gramófono	**0.851**	0.502	3.322	3.173

Tabla 5.15: Resultados promedio y desviación estándar de las métricas objetivas para el tipo de audio Vocal.

Método	ΔPEAQ	$\sigma_{\Delta\text{PEAQ}}$	ΔPAQM	$\sigma_{\Delta\text{PAQM}}$
SS Clásico	0.407	0.418	**1.839**	**1.188**
SS Denoisify	0.213	**0.357**	1.672	1.192
DL Gramófono	0.033	0.488	-1.717	3.205
DL MagTapeDB	0.414	0.535	-0.128	3.639
DL MagTapeDB + Gramófono	**0.622**	0.576	-3.021	5.720

En primer lugar, se observa que, en promedio, la mayoría de los valores de ΔPEAQ y ΔPAQM son positivos en casi todas las categorías, lo que indica que los modelos tienden a mejorar la calidad perceptual de las señales en términos

globales. Sin embargo, los valores medios difieren de manera sustancial entre los tipos de señal, y las desviaciones estándar son elevadas, lo cual sugiere que el grado de mejora no depende exclusivamente del tipo de contenido, sino también de características particulares de cada grabación.

Entre todas las categorías analizadas, la clase de *Pocas Fuentes* es la que muestra un desempeño más estable y consistente. Tanto los métodos de sustracción espectral como los modelos de aprendizaje profundo alcanzan en este grupo sus mejores resultados, especialmente según la métrica ΔPEAQ. Esto sugiere que la menor complejidad espectral de estas señales —con armónicos bien definidos y poca superposición de fuentes sonoras— facilita tanto el modelado espectral y la sustracción iterativa como la generalización de los modelos entrenados con categorías de audio diferentes.

Por otro lado, la categoría *Muchas Fuentes* representa el caso más desafiante. En términos de PEAQ, los modelos de aprendizaje profundo —particularmente aquellos entrenados con *MagTapeDB*— muestran un desempeño competitivo e incluso superior al de los métodos de sustracción espectral. Una posible explicación de este comportamiento es la presencia de música clásica y material polifónico dentro de esta categoría, cuyo contenido resulta más afín al dominio acústico de las grabaciones incluidas en *MusicNet*, utilizada durante el entrenamiento. Esta mayor cercanía entre los patrones espectrales del conjunto evaluado y los datos de entrenamiento facilita la capacidad de generalización de los modelos basados en aprendizaje profundo, lo cual se refleja en el mejor desempeño observado para este tipo de señales.

En esta categoría se observa, además, una diferencia marcada entre los dos métodos de sustracción espectral. Según los valores de ΔPEAQ, SS Denoisify obtiene un desempeño considerablemente peor que SS Clásico, lo que sugiere que el modelado espectral puede volverse contraproducente cuando la señal presenta una alta complejidad espectral, caracterizada por múltiples fuentes superpuestas y patrones armónicos difíciles de representar mediante el modelado sinusoidal. En estos casos, el enfoque más simple y directo resulta más adecuado y preserva de mejor manera la estructura original del audio.

Por otra parte, en el caso de *Música Popular*, el comportamiento muestra una mayor dependencia al tipo de modelo. Los métodos de sustracción espectral logran mejoras claras y estables, mientras que los modelos basados en aprendizaje profundo presentan, en promedio, un deterioro claro de la señal. A diferencia del grupo *Muchas Fuentes*, esta disparidad puede explicarse por la distancia entre la categoría *Música Popular* y el contenido presente en las bases de entrenamiento, ya que dicha música del conjunto personalizado no se encuentra suficientemente representada en el conjunto de datos de *MusicNet*.

La categoría *Vocal* también revela contrastes importantes. En este caso, *SS Denoisify* muestra un rendimiento inferior al de *SS Clásico* en ambas métricas, posiblemente porque su etapa de modelado espectral intenta preservar o reconstruir transitorios en un tipo de señal que, por lo general, carece de ellos. Este desajuste puede introducir artefactos indeseados provenientes de los transitorios preservados del ruido, afectando negativamente la calidad de la reconstrucción y explicando el

deterioro observado en esta categoría.

En los modelos de aprendizaje profundo ocurre un patrón similar al observado en otras categorías: PEAQ tiende a registrar mejoras o valores moderadamente positivos, mientras que PAQM identifica degradaciones significativas. Según lo observado empíricamente, esta discrepancia sugiere que estos modelos aplican una supresión del ruido más agresiva, priorizando la eliminación del ruido por sobre la preservación de la estructura armónica de la señal.

Este comportamiento puede observarse de forma clara en el modelo *DL MagTapeDB + Gramófono*, cuya dualidad resulta particularmente marcada: alcanza algunos de los valores más altos de ΔPEAQ, pero simultáneamente obtiene algunos de los peores resultados en ΔPAQM. En otras palabras, logra una reducción del ruido muy efectiva —aspecto que PEAQ tiende a valorar positivamente—, pero lo hace a costa de introducir distorsiones que alteran componentes musicales relevantes. Dado que PAQM es más sensible a la preservación tímbrica y a la fidelidad de los armónicos, penaliza con severidad estas distorsiones, lo que explica el deterioro observado en esta métrica.

Para complementar el análisis, las Figuras 5.12–5.15 presentan gráficos de dispersión que muestran la relación entre ΔPEAQ (eje x) y ΔPAQM (eje y) para cada señal procesada. Esta representación permite evaluar simultáneamente el efecto del *denoising* en ambas dimensiones perceptuales: los puntos en el cuadrante superior derecho indican mejoras conjuntas, mientras que los del cuadrante inferior izquierdo reflejan deterioro en ambas métricas.

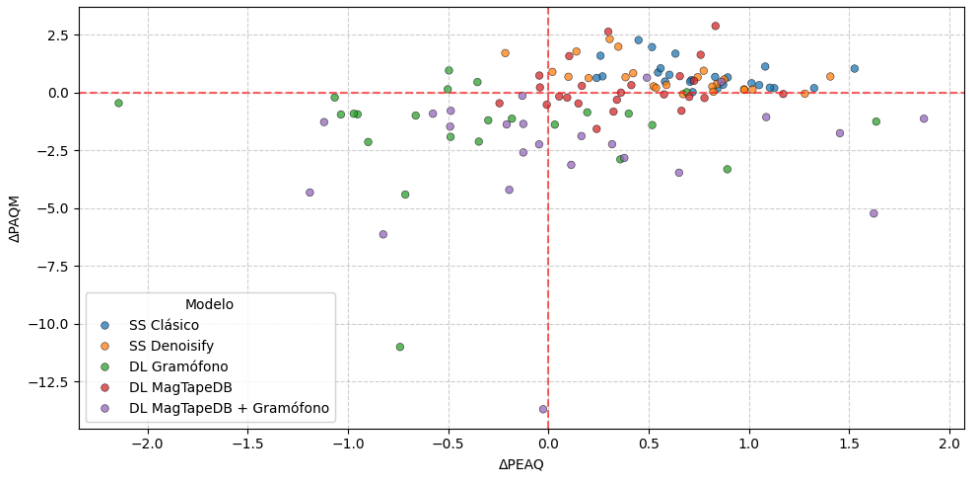

Figura 5.12: Distribución de ΔPEAQ y ΔPAQM para la categoría **Música Popular**.

De manera consistente con los resultados anteriores, los métodos de sustracción espectral exponen una distribución compacta y estable en todas las categorías. En la Figura 5.12, por ejemplo, *SS Clásico* y *SS Denoisify* se concentran mayoritariamente en la región positiva, con una dispersión moderada y muy pocos casos de degradación simultánea. Este comportamiento confirma su naturaleza robusta e independiente del tipo de contenido musical: aun cuando la categoría es exigente, el método rara vez produce distorsiones severas.

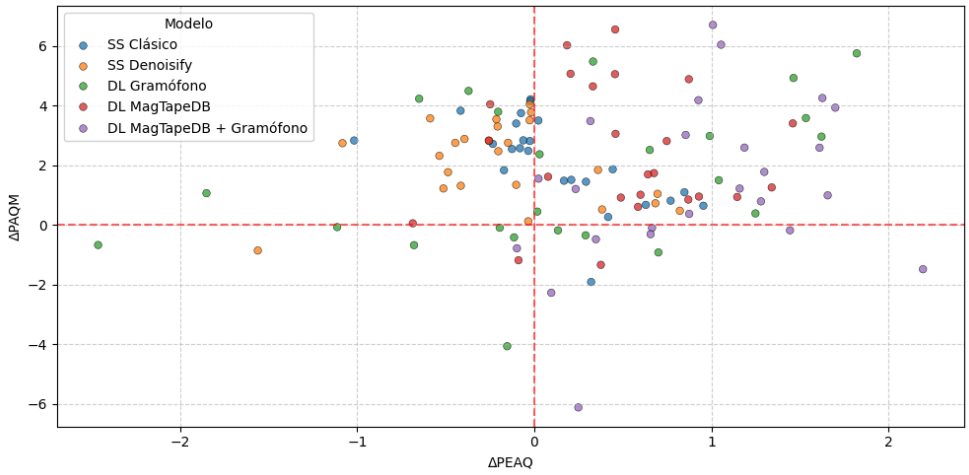

Figura 5.13: Distribución de ΔPEAQ y ΔPAQM para la categoría **Muchas Fuentes**.

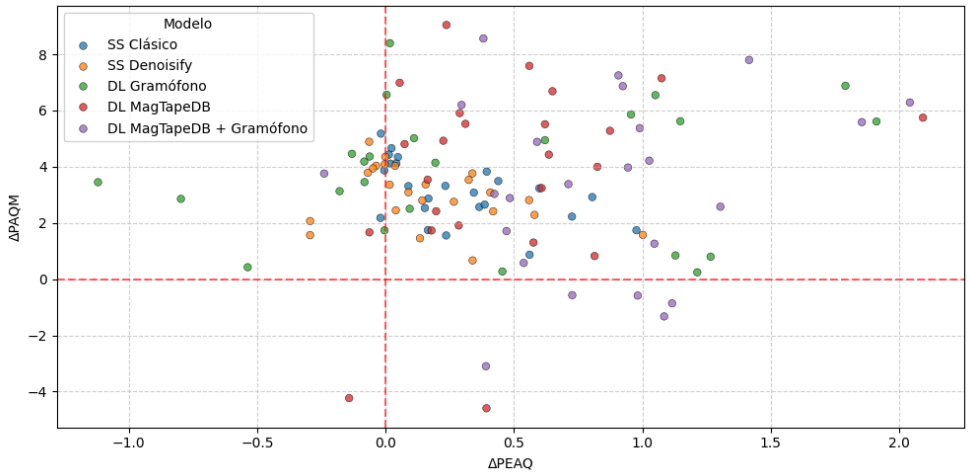

Figura 5.14: Distribución de ΔPEAQ y ΔPAQM para la categoría **Pocas Fuentes**.

Por el contrario, los modelos basados en aprendizaje profundo muestran una variabilidad notablemente mayor. Un ejemplo claro se observa también en la Figura 5.12, donde varios puntos correspondientes a *DL Gramófono* y *DL MagTapeDB + Gramófono* se sitúan en la región negativa, indicando un deterioro perceptual en ambas métricas. Sin embargo, esta tendencia cambia drásticamente en categorías más afines al contenido de sus bases de entrenamiento. En la Figura 5.13, dichos modelos pasan a registrar numerosos casos positivos —en algunos casos entre los mejores del conjunto— aunque manteniendo una dispersión considerable. Este comportamiento reafirma su capacidad de lograr mejoras significativas, pero principalmente cuando la señal de entrada se asemeja al dominio de *MusicNet*.

Los resultados evidencian que la eficacia de cada método depende fuertemente de las características del contenido espectral de la señal. Los métodos de sustracción

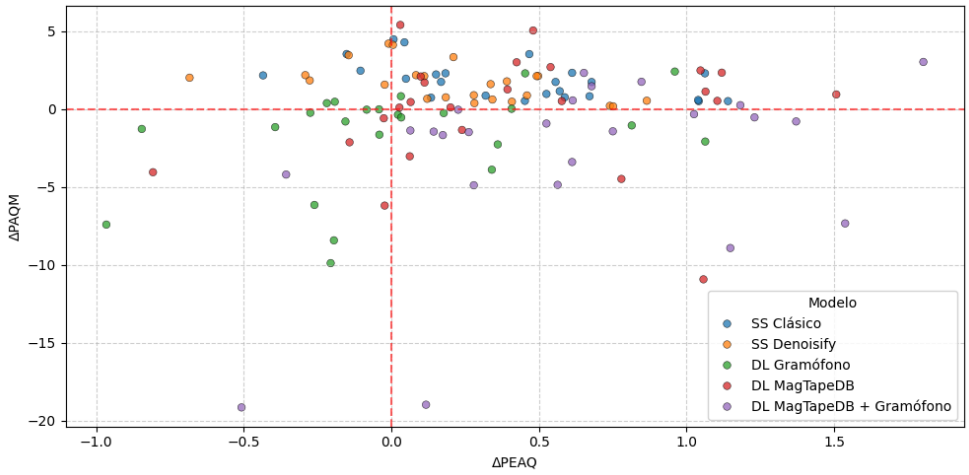

Figura 5.15: Distribución de ΔPEAQ y ΔPAQM para la categoría **Vocal**.

espectral muestran mayor estabilidad y predictibilidad entre categorías, mientras que los modelos de aprendizaje profundo exponen un desempeño más variable, fuertemente condicionado por el dominio acústico de su entrenamiento.

5.3.4. Desempeño en tiempos de ejecución

Además de las métricas perceptuales, se analizó el tiempo promedio de procesamiento de cada modelo, considerando la duración total del flujo de inferencia o del algoritmo correspondiente. La Figura 5.16 muestra los tiempos medios y su dispersión, mientras que la Tabla 5.16 resume los valores numéricos obtenidos.

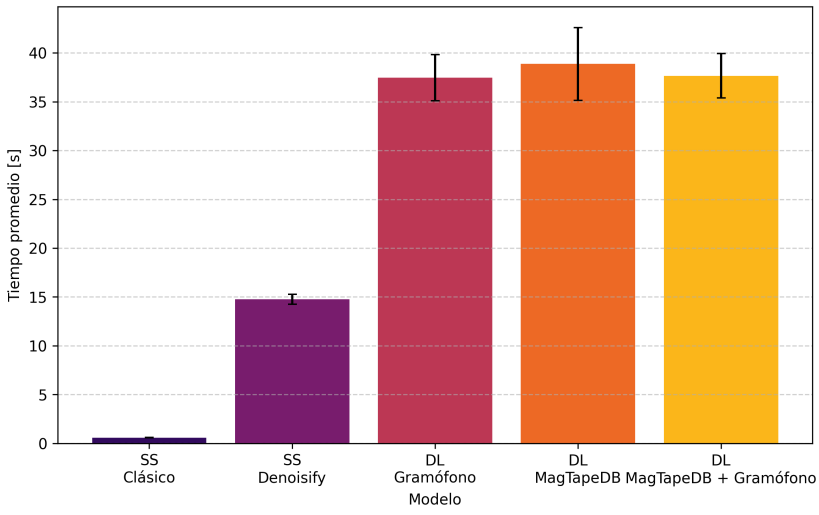

Figura 5.16: Tiempos promedio de procesamiento por modelo. La figura evidencia las diferencias de demanda computacional entre técnicas de sustracción espectral y modelos de aprendizaje profundo.

Tabla 5.16: Tiempos promedio de procesamiento por modelo, con su desviación estándar.

Método	**Tiempo medio [s]**	**Desv. estándar [s]**
SS Clásico	**0.592**	**0.020**
SS Denoisify	14.774	0.509
DL Gramófono	37.462	2.378
DL MagTapeDB	38.861	3.743
DL MagTapeDB + Gramófono	37.650	2.276

Los resultados evidencian una diferencia significativa en la demanda computacional entre los métodos basados en procesamiento de señales y los modelos de aprendizaje profundo. El algoritmo *SS Clásico* presenta el menor tiempo promedio de ejecución (alrededor de 0.6 s por archivo), seguido por *SS Denoisify*, cuyo tiempo medio asciende a unos 15 s. En este último caso, el incremento se debe principalmente a la incorporación de modelado espectral y de la sustracción iterativa, cuyo tiempo total depende del número de iteraciones configuradas (en este caso 30). Ambos métodos son totalmente ejecutables en CPU sin necesidad de aceleración por GPU, lo que refuerza su aplicabilidad en contextos de bajo costo computacional o en entornos de procesamiento en tiempo real.

Por otra parte, los modelos de aprendizaje profundo presentan tiempos de inferencia considerablemente mayores, en torno a 37–39 s por archivo. Aunque tampoco requieren GPU para la inferencia, su ejecución se beneficia notablemente de su uso, como en este caso. Además, debe considerarse que estos valores no reflejan los requisitos computacionales del entrenamiento de las redes neuronales, los cuales son órdenes de magnitud superiores tanto en tiempo como en demanda de recursos, y constituyen una etapa sustancial en el desarrollo de estos modelos.

En total, los resultados de esta sección muestran que los métodos de sustracción espectral —tanto el clásico como su versión con modelado espectral— ofrecen una buena relación entre desempeño perceptual y eficiencia computacional. Si bien los modelos de aprendizaje profundo logran en algunos casos resultados perceptualmente superiores, los métodos basados en procesamiento de señales alcanzan rendimientos comparables, e incluso mejores en ciertos contextos, con tiempos de procesamiento mucho menores. Esto sugiere que, al menos dentro del alcance de este experimento, las soluciones de procesamiento clásico continúan siendo una alternativa altamente competitiva en tareas de restauración de audio, especialmente cuando la eficiencia es un factor determinante.

En la siguiente sección se complementa este análisis con una evaluación subjetiva de las grabaciones procesadas, contrastando las observaciones perceptuales con los resultados cuantitativos obtenidos en las métricas objetivas.

5.4. Escucha crítica de las señales restauradas

En esta sección se presentan los resultados de la escucha crítica realizada sobre las grabaciones restauradas. No se trata de un análisis exhaustivo, sino de la

selección de algunas grabaciones representativas que permiten ilustrar ciertos problemas observados y valorar de manera cualitativa el desempeño de las técnicas evaluadas.

En términos generales, ambas técnicas —tanto la sustracción espectral como el modelo de aprendizaje profundo— producen señales con una reducción de ruido claramente perceptible y, en la mayoría de los casos, una preservación adecuada de la señal, obteniendo un resultado final satisfactorio. No obstante, con el fin de complementar el análisis objetivo, se describen aquí las principales distorsiones detectadas en los casos donde estas técnicas presentan limitaciones.

Se realizará un análisis cualitativo de las distorsiones introducidas por los modelos *DL MagTapeDB* y *SS Clásico / SS Denoisify*, dado que fueron los métodos que mostraron el mejor desempeño global en las evaluaciones objetivas. Para este fin se seleccionó un ejemplo en el que las distorsiones resultan particularmente evidentes, lo cual permite ilustrar con claridad los límites y comportamientos característicos de cada enfoque.

A fin de mantener la comparación en un escenario realista, se utilizaron señales con una SNR de 16 dB, valor que se aproxima al nivel de ruido típico observado en las grabaciones históricas de Lauro Ayestarán, según el trabajo de I.Irigaray *et al.* [5].

Es importante destacar que las distorsiones aquí presentadas no aparecen de forma sistemática en todas las señales: su presencia y magnitud varían según el contenido musical y las particularidades de cada audio. El ejemplo seleccionado corresponde, por tanto, a un caso representativo pero deliberadamente exigente que permite examinar con mayor claridad los artefactos generados por cada modelo.

Finalmente, se incluye también un análisis auditivo sobre fragmentos reales del archivo sonoro de la colección de Lauro Ayestarán, procesados con los tres modelos considerados. Estos ejemplos permiten observar cómo se trasladan los comportamientos identificados en señales sintéticas o controladas a un escenario histórico y acústicamente más complejo. Además, para acompañar el análisis y facilitar la exploración de los resultados obtenidos, se desarrolló una página web interactiva [44].

5.4.1. Distorsiones resultantes de la restauración

Ruido tonal y agudo

La primera distorsión analizada aparece al procesar una balada interpretada por piano y voz femenina. En este caso, la distorsión producida por el modelo *DL MagTapeDB* se manifiesta como un tono agudo en torno a los 6 kHz, claramente audible aunque de impacto menor en comparación con otras distorsiones que se describirán más adelante. En la Figura 5.17 se muestran los espectrogramas correspondientes a este ejemplo, en el siguiente orden: audio limpio, audio contaminado con ruido a 16 dB, audio restaurado y residuo. En el tercer espectrograma se aprecia con claridad la aparición del tono en cuestión, visible como una línea horizontal localizada aproximadamente entre 5 y 6 kHz, ausente en la señal original.

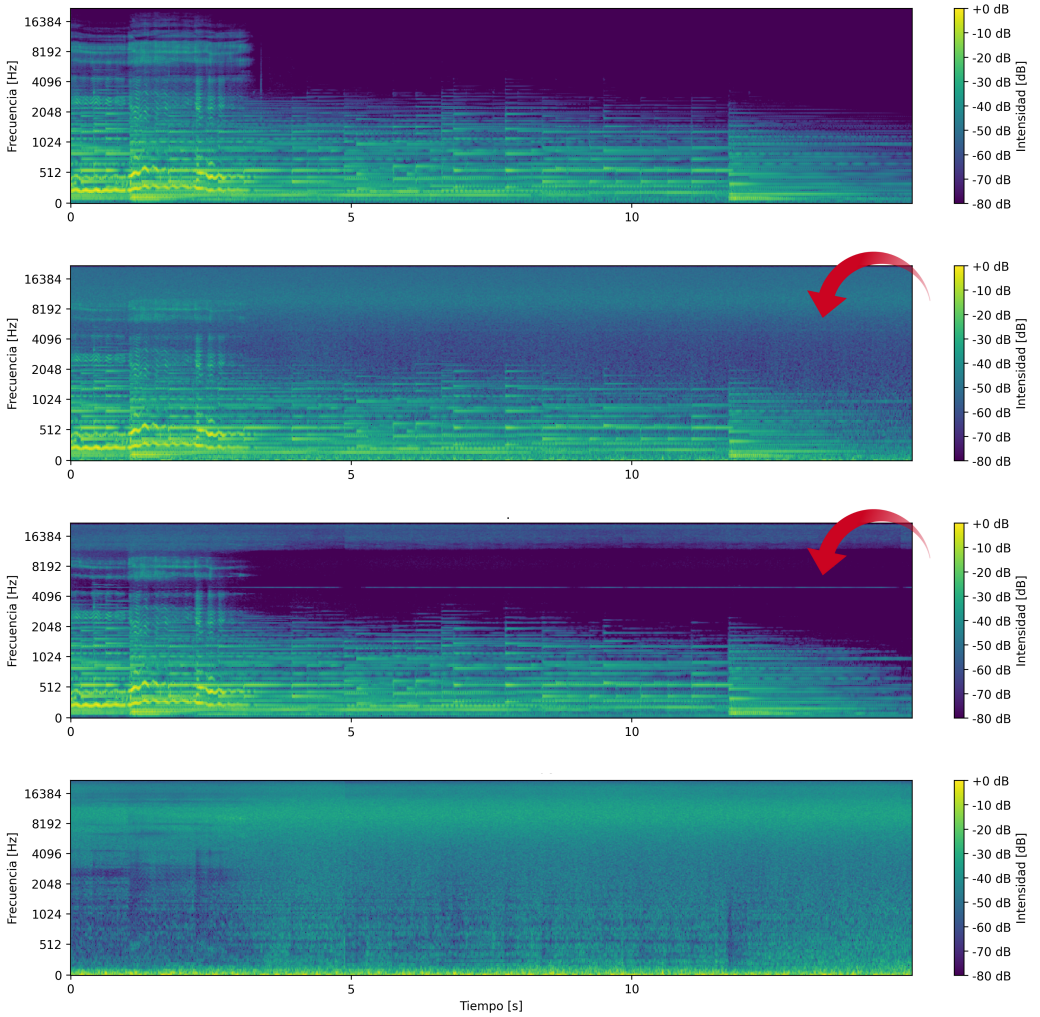

Figura 5.17: Espectrogramas del ejemplo seleccionado: audio limpio, audio contaminado con ruido de cinta a 16 dB, señal restaurada mediante *DL MagTapeDB* y residuo. En la señal restaurada se aprecia un tono agudo artificial —marcado en **rojo**— alrededor de 5–6 kHz, producto del realce involuntario de un componente del ruido que el modelo interpreta erróneamente como parte de la señal original.

Esto ocurre porque el ruido original contiene una componente marcada en esa banda, y el modelo atenúa fuertemente el contenido circundante, dejando dicha componente aislada. De este modo, no se trata de un artefacto generado desde cero por el modelo, sino de un componente del propio ruido que es interpretado erróneamente como parte de la señal útil. Esto también explica su presencia intermitente a lo largo del conjunto de evaluaciones: bajo otras condiciones puede quedar enmascarado o, directamente, no aparecer en el ruido específico utilizado.

Este fenómeno ilustra nuevamente una limitación estructural de los modelos de aprendizaje profundo aplicados a la restauración de audio: al basar sus decisiones en patrones estadísticos aprendidos durante el entrenamiento, pueden realzar

inadvertidamente componentes del ruido que coinciden con dichas regularidades, generando así distorsiones nuevas o amplificando elementos espurios presentes en la señal de entrada.

Los valores correspondientes a esta restauración se presentan en la Tabla 5.17, donde se observa el desempeño del modelo frente al ejemplo analizado en comparación con sus promedios globales y con los promedios obtenidos dentro de la categoría *Pocas Fuentes*.

Tabla 5.17: Valores de ΔPEAQ y ΔPAQM obtenidos por el modelo *DL MagTapeDB* al restaurar un ejemplo con ruido de cinta a 16 dB. Se incluyen, a modo de referencia, los promedios: global y correspondiente a la categoría *Pocas fuentes*.

Referencia	**ΔPEAQ**	**ΔPAQM**
Ejemplo (16 dB)	0.311	5.525
Promedio Pocas Fuentes	0.481	3.807
Promedio Global	0.439	1.575

Vale la pena destacar que el valor de ΔPEAQ obtenido es inferior tanto al promedio global del modelo como al promedio específico de la categoría *Pocas Fuentes*, lo que indica que la restauración resulta perceptualmente menos efectiva que en la mayoría de los casos evaluados. No obstante, el valor de ΔPAQM supera ampliamente ambos promedios, reflejando que, pese a la presencia de distorsión tonal, el método logra una mejora sustancial desde la perspectiva de esta métrica.

Filtrado de bajas frecuencias

El filtrado de bajas frecuencias esta presente en varios audios de la base de datos para el modelo *DL MagTapeDB*, aunque en la mayoría de los casos con menor impacto perceptual. Sin embargo, en el ejemplo que se presenta a continuación, la distorsión adquiere un carácter severo y modifica de forma significativa la estructura espectral de la señal.

La distorsión analizada consiste en una atenuación pronunciada de las bajas frecuencias por parte del modelo, fenómeno claramente visible en la Figura 5.18. En el espectrograma del audio limpio se aprecian ataques percutivos con alta energía en las bandas graves; estos ataques siguen presentes tras la adición de ruido, pero resultan notablemente reducidos luego del proceso de *denoising*. El residuo —cuarto espectrograma— concentra gran parte de la energía eliminada justamente en esas zonas de baja frecuencia, evidenciando que el modelo suprime componentes legítimos de la señal. Este efecto también se percibe, aunque de forma más sutil, en el audio restaurado, donde la región grave aparece visiblemente empobrecida respecto del audio original.

El fragmento considerado posee una instrumentación particularmente diversa, con timbres poco habituales y una mezcla compleja. Podría suponerse que la ausencia de varios de estos instrumentos en la base de entrenamiento conduciría a una eliminación más agresiva de los mismos; sin embargo, el modelo no atenúa de forma significativa estos timbres inusuales, sino que afecta principalmente los elementos

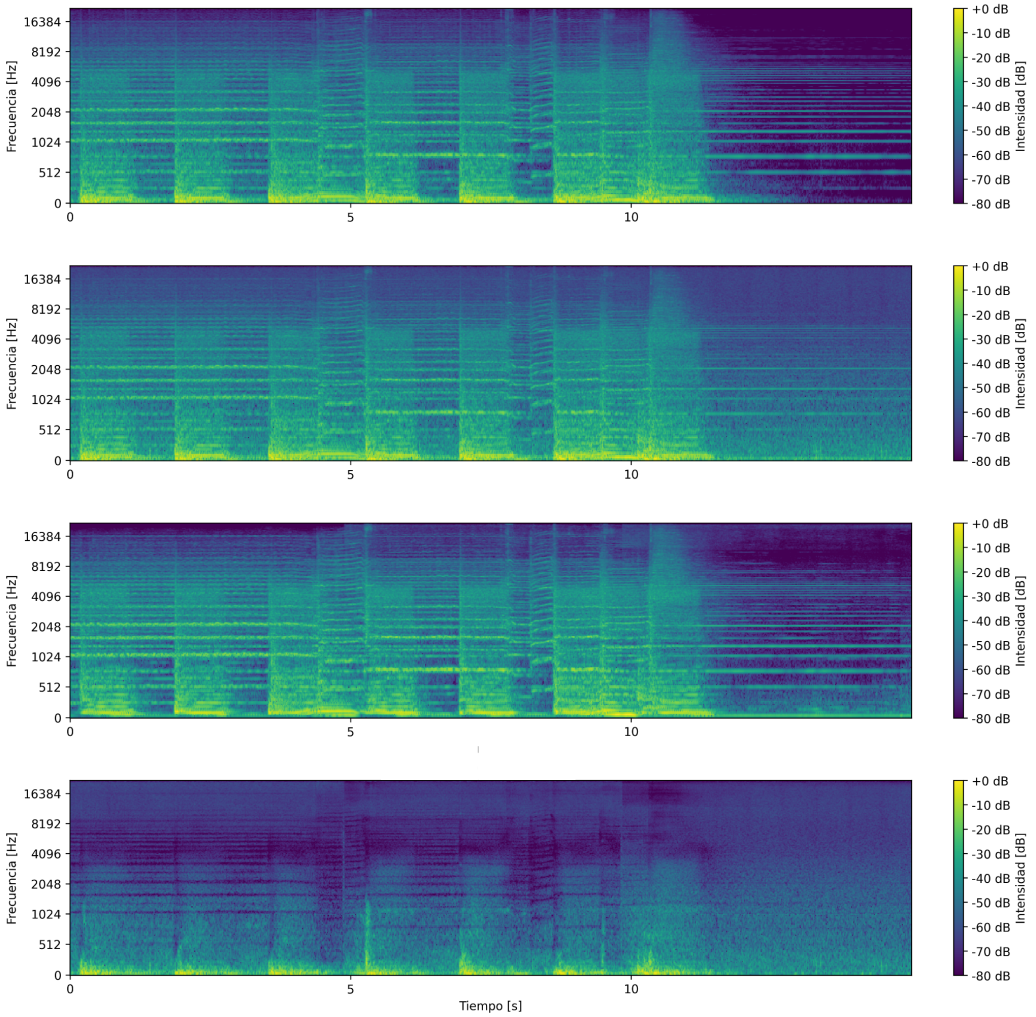

Figura 5.18: Espectrogramas del ejemplo analizado: audio limpio, audio contaminado con ruido de cinta a 16 dB, audio restaurado con *DL MagTapeDB* y residuo, respectivamente. En este último puede observarse la atenuación excesiva de componentes de baja frecuencia introducida por el método.

percutivos de la mezcla, aun cuando dicho tipo de contenido está representado en los datos de entrenamiento.

Este comportamiento sugiere que la distorsión observada no se explica únicamente por la falta de familiaridad con ciertos timbres, sino por la forma en que el modelo interpreta la energía transitoria en bajas frecuencias, tendiendo a confundirla con ruido y suprimiéndola en exceso.

En este caso, las métricas objetivas presentadas en la Tabla 5.18 evalúan la restauración de forma marcadamente negativa. Ambas métricas muestran valores muy inferiores a sus respectivos promedios —e incluso negativos en el caso de ΔPAQM— lo que indica que el proceso de *denoising* no solo no mejora la señal, sino que la degrada perceptualmente en comparación con el audio ruidoso. Esta

penalización severa es coherente con la distorsión analizada anteriormente, donde el modelo atenúa de manera excesiva las componentes de baja frecuencia, eliminando información estructural relevante del audio original.

Tabla 5.18: Valores de ΔPEAQ y ΔPAQM obtenidos por el modelo *DL MagTapeDB* al restaurar un ejemplo contaminado con ruido de cinta a 16 dB. Se incluyen, como referencia, los promedios globales del modelo y los promedios correspondientes a la categoría *Muchas fuentes*.

Referencia	**ΔPEAQ**	**ΔPAQM**
Ejemplo (16 dB)	0.376	-1.337
Promedio Muchas Fuentes	0.484	2.393
Promedio Global	0.439	1.575

Eliminación de transitorios

Esta distorsión es la más frecuente identificada en la base de datos de música personalizada y aparece en todos los modelos evaluados. En el caso del modelo *DL MagTapeDB*, su impacto es particularmente severo, generando una sensación perceptual de audio "ahogado" debido a la supresión sistemática de transitorios.

Un ejemplo representativo se muestra en la Figura 5.19, donde se ilustran los segundos finales de un audio de folklore uruguayo interpretado con guitarra y voz. En el espectrograma del audio restaurado, las regiones marcadas en rojo corresponden a transitorios que son eliminados por completo, mientras que otros son atenuados parcialmente.

Esta pérdida se refleja en el espectrograma del residuo, donde aparecen barras verticales de alta energía asociadas a estos eventos transitorios descartados. Dado que estructuras de este tipo no están adecuadamente representadas en el conjunto de entrenamiento del modelo, su supresión resulta consistente con lo observado en otros audios con características similares.

Aunque de manera menos pronunciada, los métodos de sustracción espectral también presentan pérdidas parciales de transitorios. En la Figura 5.20 se muestra el mismo fragmento procesado mediante *SS Denoisify*. A diferencia del modelo de aprendizaje profundo, el transitorio marcado en rojo se preserva, aunque con menor intensidad. El residuo revela líneas verticales que confirman una atenuación generalizada de estos eventos.

Además de la pérdida de transitorios, ambas figuras permiten identificar otras distorsiones relevantes. Las regiones marcadas en violeta muestran la supresión completa de componentes agudas que, si bien estaban fuertemente enmascaradas por el ruido, forman parte de la señal original. Un fenómeno relacionado ocurre en el decaimiento final del audio: al encontrarse parcialmente oculto por el ruido, el modelo lo interpreta como ruido residual y lo elimina, generando un final perceptualmente más abrupto que el presente en la grabación limpia.

A pesar de la presencia de las distorsiones descritas, las métricas objetivas confirman que, en ambos modelos, el resultado restaurado constituye una mejora respecto al audio ruidoso, tal como se aprecia en las Tabla 5.19 y Tabla 5.20.

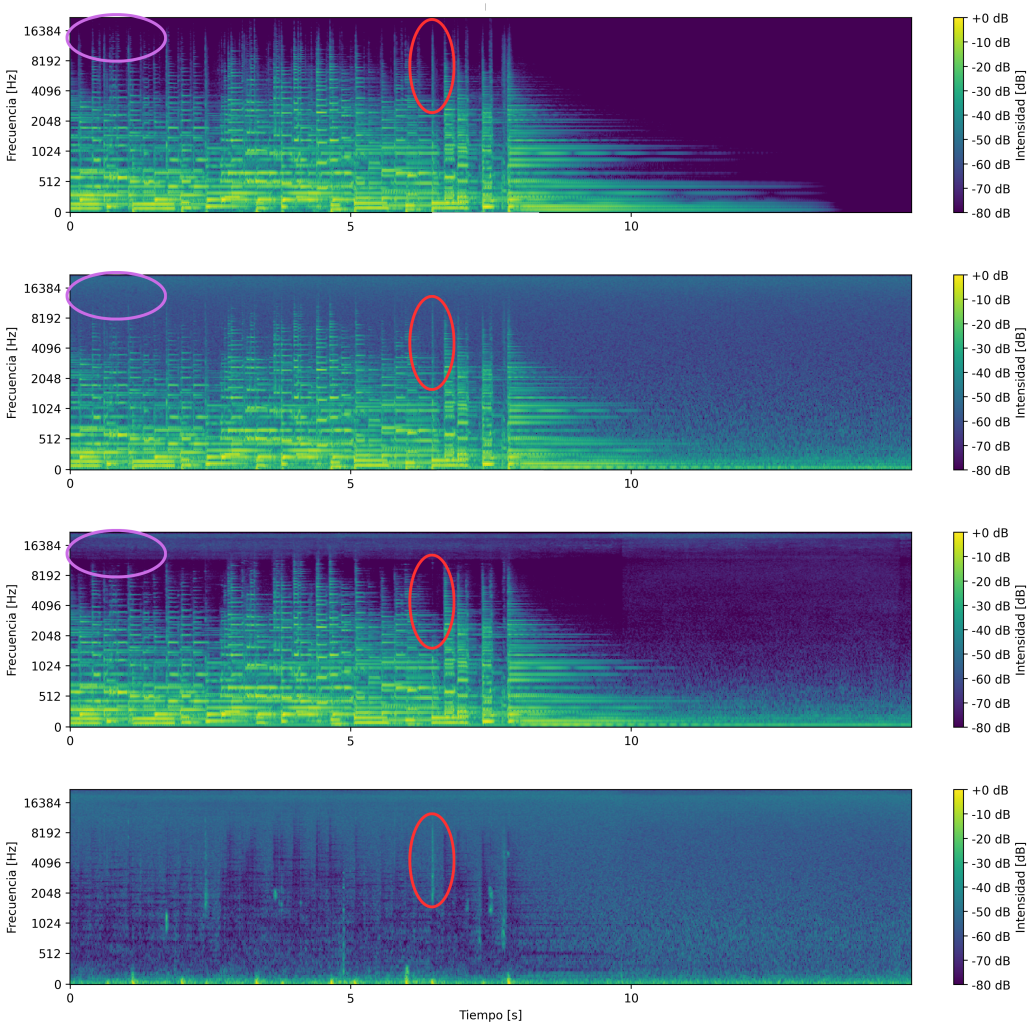

Figura 5.19: Espectrogramas de un pasaje de guitarra y voz: audio limpio, audio contaminado con ruido de cinta a 16 dB, audio restaurado con DL MagTapeDB y residuo, respectivamente. En **rojo** se indican transitorios eliminados y en **violeta** la supresión de componentes agudas enmascaradas por el ruido.

En el caso de *DL MagTapeDB*, el desempeño obtenido para este ejemplo se sitúa por encima tanto del promedio global del modelo como del promedio correspondiente a la categoría *Pocas fuentes*. Esto indica que, si bien el audio restaurado presenta artefactos perceptibles, no se trata de uno de los casos más severos dentro del conjunto evaluado; por el contrario, mantiene un rendimiento claramente superior al promedio. Algo similar ocurre con *SS Denoisify*, cuyos valores también se mantienen por encima del promedio global, aun cuando la distorsión de transitorios sigue siendo apreciable en la señal restaurada.

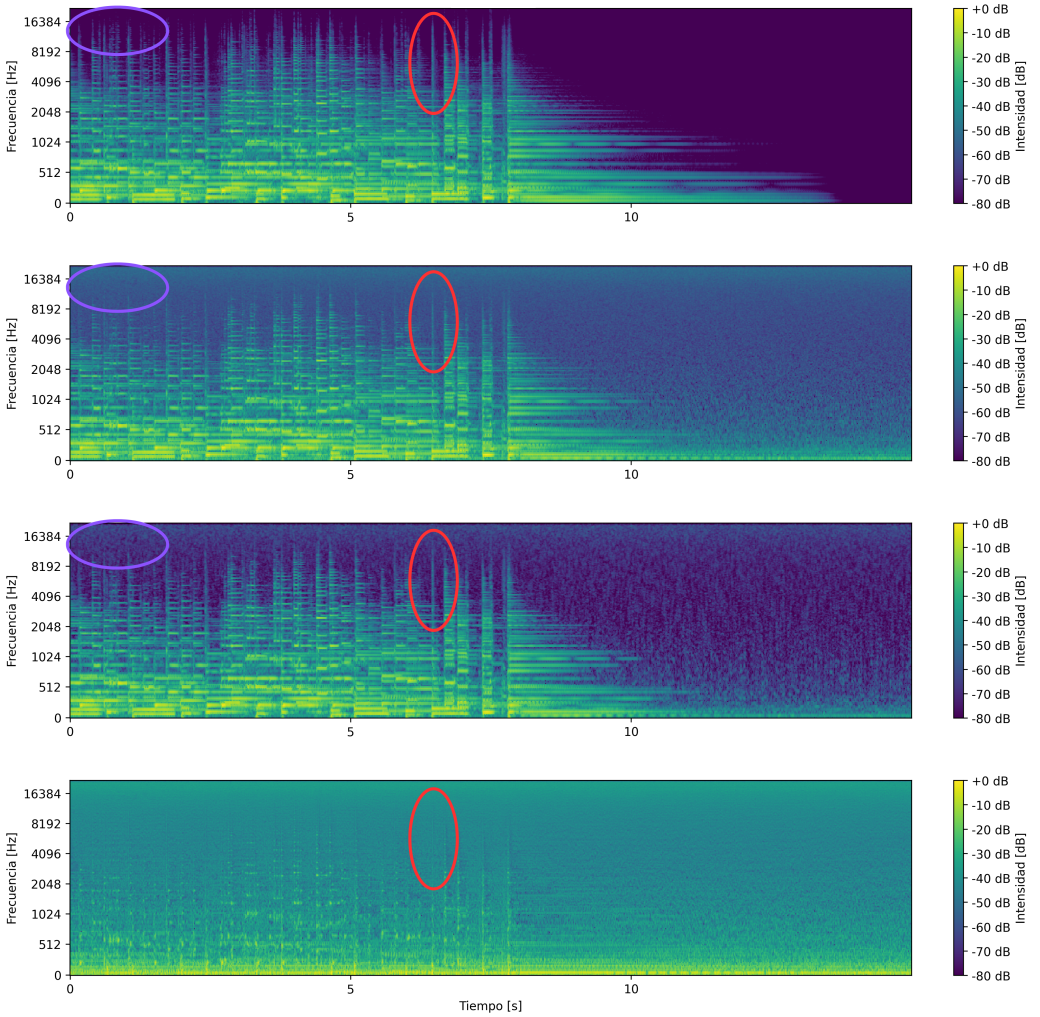

Figura 5.20: Espectrogramas del mismo fragmento procesado con SS Denoisify. En **rojo** se preserva un transitorio; en **violeta** se observan componentes agudas suprimidas; el tercer espectrograma muestra ruido musical introducido por el método.

Tabla 5.19: Tabla de $\Delta PEAQ$ y $\Delta PAQM$ para el tema *Milagro* (Larbanois - Carrero) eliminando ruido de SNR 10dB y 16dB a travez del método DL MagTapeDB, en comparación con valores promedio

Referencia	$\Delta PEAQ$	$\Delta PAQM$
Ejemplo (16dB)	0.608	3.242
Promedio Pocas Fuentes	0.481	3.807
Promedio Global	0.439	1.575

Ruido musical

La distorsión predominante en los modelos de sustracción espectral es el *ruido musical*, presente en todos los audios de la base de datos con distintos grados de

Tabla 5.20: Valores de ΔPEAQ y ΔPAQM obtenidos por el método *SS Denoisify* al restaurar un ejemplo con ruido de cinta a 16 dB. Se incluyen, como referencia, los promedios globales del modelo y los correspondientes a la categoría *Pocas fuentes*.

Referencia	**ΔPEAQ**	**ΔPAQM**
Ejemplo (16 dB)	0.266	2.755
Promedio Pocas Fuentes	0.168	3.007
Promedio Global	0.202	1.856

intensidad.

Tal como se observa en la Figura 5.20, el ruido musical se identifica con claridad al comparar las regiones de silencio entre los distintos espectrogramas. En el audio limpio, estas zonas aparecen como áreas lisas y uniformes, de color azul oscuro, reflejando la ausencia de energía. Tras la adición de ruido, dichas regiones se aclaran y adoptan un tono celeste y verde, coherente con el incremento de energía de banda ancha, aunque aún conservan un patrón visual relativamente homogéneo. En el audio restaurado, si bien la energía del ruido disminuye —visible por la tonalidad más oscura—, la textura se vuelve granular, fragmentada y no uniforme, lo que constituye la firma típica del ruido musical.

Tabla 5.21: Valores de ΔPEAQ y ΔPAQM obtenidos por el método *SS Clásico* al restaurar un ejemplo con ruido de cinta a 16 dB. Se incluyen, como referencia, los promedios globales del modelo y los correspondientes a la categoría *Pocas fuentes*.

Referencia	**ΔPEAQ**	**ΔPAQM**
Ejemplo (16 dB)	0.386	2.651
Promedio Pocas Fuentes	0.281	3.117
Promedio Global	0.392	1.963

Los valores objetivos correspondientes se presentan en la Tabla 5.20. Aunque la restauración recibe una calificación positiva —indicando una mejora respecto al audio contaminado—, el desempeño se sitúa por debajo de los promedios globales y específicos de la categoría. Esto sugiere que, en este ejemplo en particular, la presencia de ruido musical es más intensa que en otros casos del conjunto evaluado.

Como era de esperar, el audio procesado mediante *SS Clásico* presenta un comportamiento muy similar. La tabla incluida en la Tabla 5.21 confirma esta cercanía en el rendimiento, lo que pone de manifiesto que ambos métodos comparten las mismas limitaciones estructurales inherentes a la sustracción espectral.

5.4.2. Análisis sobre grabaciones de archivo musical

Con el fin de analizar el comportamiento de los algoritmos frente a material histórico real, se presentan a continuación dos estudios de caso basados en grabaciones auténticas, es decir, registros cuyo ruido no ha sido agregado sintéticamente. Al tratarse de materiales sin versión "limpia", no es posible aplicar las métricas

objetivas previamente utilizadas. No obstante, resulta pertinente realizar una escucha crítica y un análisis cualitativo de los resultados obtenidos con los tres métodos con mejor desempeño general: los dos enfoques de sustracción espectral y el modelo entrenado con la base *MagTapeDB*.

Caso 1: "Estilo", por Amalia de la Vega

Para este análisis se seleccionó una grabación histórica de Amalia de la Vega perteneciente al archivo de Lauro Ayestarán. Tal como documenta Ruiz [45], y citando textualmente:

> "El 19 de marzo de 1949 Amalia de la Vega grabó para Ayestarán en una sesión hecha en la casa del musicólogo, en ese entonces ubicada en la calle Chuy 3208, en Montevideo. Tenía 35 años e interpretó cinco estilos, cinco milongas, dos cifras y una vidalita. Se trata de las únicas grabaciones conocidas en las que se acompaña a sí misma con guitarra." [45]

Este registro corresponde a 30 segundos de uno de los estilos[1] interpretados por la artista en dicha sesión. En la Figura 5.21 se muestra una fotografía de Amalia de la Vega durante la década en que fue registrada por Ayestarán.

Figura 5.21: Amalia de la Vega, quien en 1949 realizó para Ayestarán una sesión de grabación en la que se acompañó a sí misma con guitarra.

Es importante aclarar que la fuente original no es una cinta, sino un disco instantáneo de 78 rpm, que posteriormente fue respaldado en cinta magnética. Por lo tanto, el ruido presente en el material actual es, principalmente, la suma de dos degradaciones distintas: el ruido de superficie característico de los discos instantáneos y el propio de la cinta magnética. (Ver Anexo B.1 y Anexo B.2).

En 1992, Walter Díaz realizó una transferencia en casetes C90 de las cintas y posteriormente, se presume en 1993, se efectuó una nueva copia de esos casetes hacia cinta de audio digital.

[1] El Estilo es un género musical folclórico rioplatense caracterizado por el canto acompañado de guitarra. También se lo conoce como Triste, por su carácter melancólico, o Décima, debido a su estructura poética de diez versos. [46].

En consecuencia, el archivo que llega a la actualidad no solo hereda los ruidos propios del disco instantáneo original y de la cinta analógica intermedia, sino que también incorpora el ruido añadido en la copia a casete y los posibles artefactos menores de la digitalización, como el ruido de cuantización.

En consecuencia, el archivo que llega a la actualidad no solo hereda los ruidos propios del disco instantáneo original y de la cinta analógica intermedia, sino que también incorpora el ruido añadido en la copia a casete y los posibles artefactos menores de la digitalización, como el ruido de cuantización.

La Figura 5.22 ilustra el disco instantáneo de acetato utilizado en la sesión.

Figura 5.22: Disco de acetato de base metálica de 25 cm de diámetro utilizado por Lauro Ayestarán para grabar a Amalia de la Vega en 1949. Archivo del CDM.

En el audio original se perciben con claridad la voz de Amalia de la Vega y el acompañamiento de guitarra, ambos inmersos en un ruido de fondo. Aunque parte de ese ruido proviene del hiss de la cinta, la textura granular e irregular podría estar relacionada con el disco de 78 rpm.

En el espectrograma del audio original (Fig. 5.23) se observa que el ruido concentra buena parte de su energía en las bandas altas, por encima de aproximadamente 4 kHz. Esta energía no es completamente uniforme: existen regiones donde el ruido es más denso y otras donde disminuye notablemente. El ruido presenta una textura marcadamente irregular, con aspecto estriado o rasgado. Esto podría deberse a las microimperfecciones del surco del disco original, que generan fluctuaciones finas en las altas frecuencias.

La voz y la guitarra, aunque discernibles y con presencia tímbrica definida, no están libres de distorsión, lo que produce un efecto levemente apagado o "ahogado". En este caso es importante remarcar que dicha distorsión es inherente al registro original y no un artefacto introducido por los métodos de reducción de ruido evaluados.

Por otro lado, en la voz de Amalia se perciben pequeñas oscilaciones que podrían sugerir la presencia de *dropouts*. Sin embargo, dado el origen en disco instantáneo, es igualmente plausible que se trate de ligeras variaciones mecánicas propias del soporte (pérdida momentánea de contacto, rugosidad superficial, etc.), por lo que no es posible confirmarlo con certeza en el espectrograma.

Si bien los *dropouts* pueden mitigarse parcialmente, no constituyen ruido aditivo y por ende no son el objetivo de este trabajo ni de las técnicas utilizadas. Su tratamiento requiere métodos alternativos, como técnicas de *inpainting* [47] que reconstruyen la región faltante a partir del contexto.

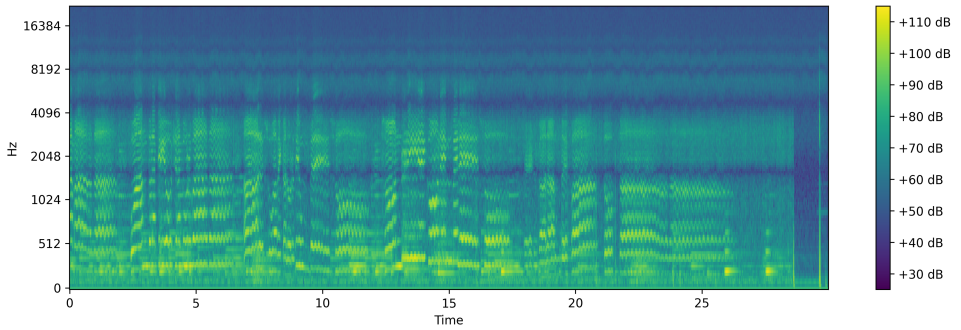

Figura 5.23: Espectrograma del audio original de Amalia de la Vega, donde se observa la distribución espectral del ruido y la presencia conjunta de la voz y la guitarra antes de cualquier proceso de reducción de ruido.

En el resultado obtenido a partir de *SS Clásico*, tanto en la escucha como en el espectrograma procesado se aprecia una reducción de ruido claramente perceptible. El método atenúa de manera efectiva el soplido de fondo, Si bien algunas de las bandas descritas previamente permanecen presentes, el efecto de *denoising* es evidente y contribuye a una mayor limpieza general del registro.

Al tener menos ruido, la guitarra se distingue con mayor claridad y su textura es muy similar a la del material original, pese a la degradación propia ya mencionada. En la Fig. 5.24 puede apreciarse que el método atenúa principalmente las componentes de ruido en las bandas altas, particularmente por encima de los 4 kHz.

No obstante, al analizar el espectrograma del residuo y escuchar el audio correspondiente, se advierte que también se elimina una cantidad muy pequeña de contenido de baja frecuencia asociado a la guitarra, aunque se requiere una escucha cuidadosa para percibirlo.

Por otra parte, la voz aparece de forma marcada en el residuo, lo que indica que parte de su energía se solapa espectralmente con el ruido y es parcialmente retirada junto con él.

En el caso de *SS Denoisify* se optó por hacer una reducción de ruido más agresiva que la utilizada en la versión clásica, ajustando algunos de los hiperparámetros por defecto del método Denoisify a fin de obtener una mayor atenuación del ruido. El resultado es coherente con lo esperado: se percibe una reducción de ruido más profunda, pero acompañada de la aparición de ruido musical.

En el espectrograma correspondiente (Fig. 5.25) este fenómeno se ve como pequeños puntos o trazas aisladas de energía, especialmente en las bandas altas donde originalmente predominaba el ruido.

El residuo de este método muestra, efectivamente, una mayor cantidad de contenido removido en comparación con la versión clásica, incluyendo tanto más

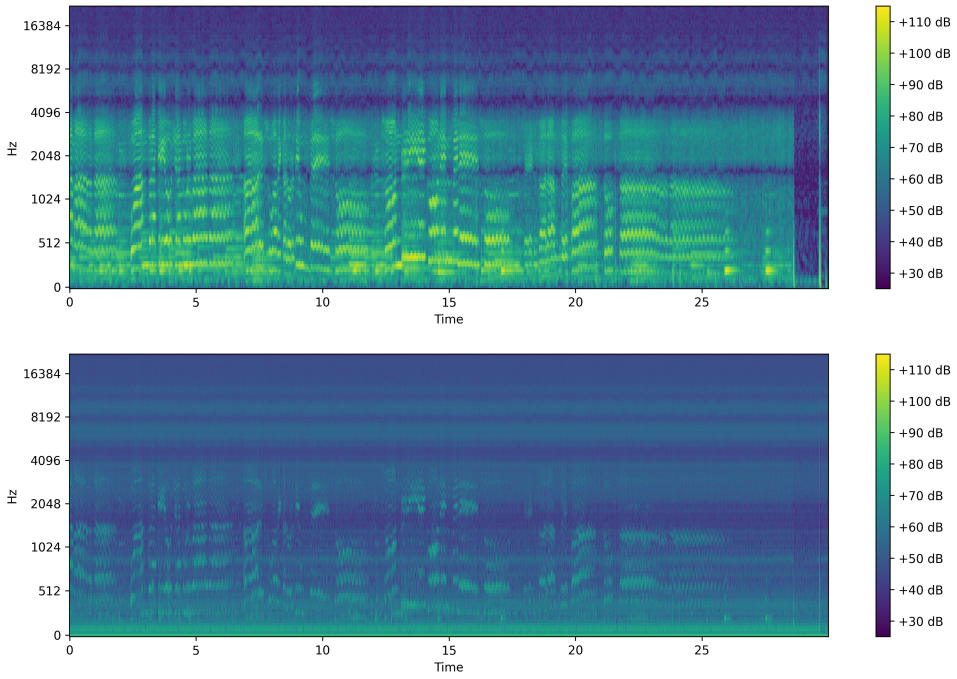

Figura 5.24: Espectrogramas del método **SS Clásico** aplicado al audio de Amalia de la Vega. Arriba: señal procesada, donde se observa la atenuación del ruido de alta frecuencia y la mejora general de la claridad. Abajo: espectrograma del residuo, que muestra las componentes ruidosas eliminadas y la pequeña porción de señal útil retirada.

ruido como una pequeña fracción adicional de señal útil, lo cual coincide con la impresión auditiva.

Como es habitual en este tipo de técnicas, se observa nuevamente el compromiso entre una mayor reducción de ruido y el incremento de artefactos perceptuales. En este caso particular, el método puede permitirse ser más agresivo debido a que el registro original ya presenta una degradación inherente en la voz y la guitarra; es decir, la riqueza tímbrica del material no es especialmente alta. No obstante, el grado de agresividad adecuado continúa siendo una decisión subjetiva y depende del criterio y las prioridades de cada oyente.

En términos globales, los resultados obtenidos con este método son similares a los de la sustracción espectral clásica, salvo por las diferencias ya mencionadas. No obstante, una ventaja importante de esta variante radica en la flexibilidad que ofrece a través de los hiperparámetros descritos anteriormente en la Sección 2.6. Dependiendo del aspecto del audio que se desee priorizar, es posible ajustar el comportamiento del algoritmo para obtener un resultado más centrado en las componentes tonales de la voz o, alternativamente, para preservar en mayor medida los transitorios y el carácter rítmico de la guitarra.

El método basado en aprendizaje profundo (*DL MagTapeDB*) presenta un comportamiento claramente diferenciado respecto a las variantes de sustracción espectral. En primer lugar, se observa una atenuación mucho más agresiva en las

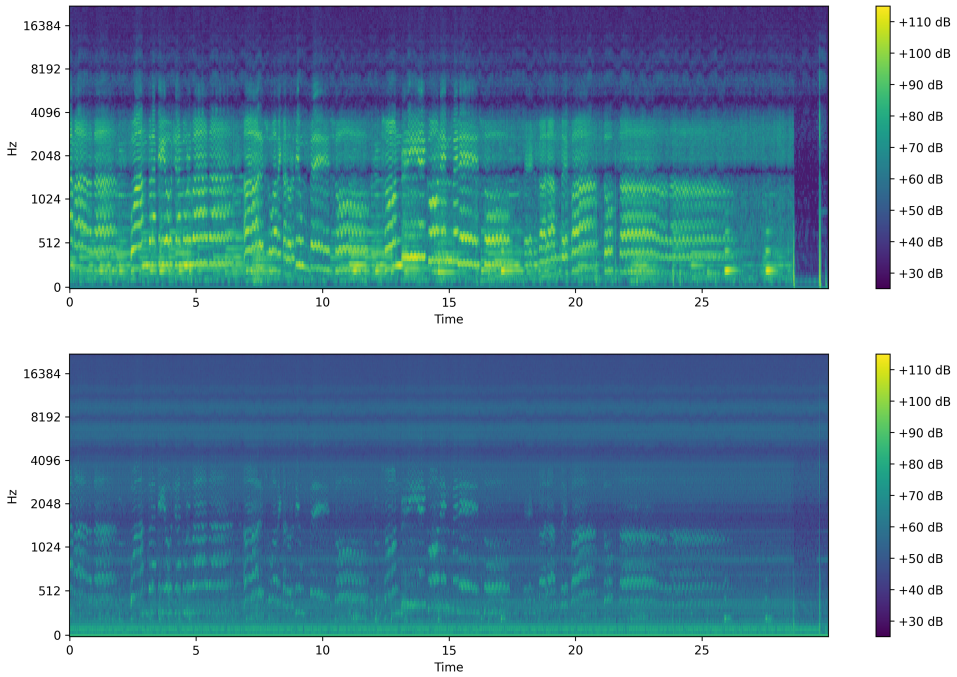

Figura 5.25: Espectrogramas del método **SS Denoisify**. Arriba: señal procesada, donde se aprecia una reducción de ruido más agresiva acompañada de ruido musical. Abajo: residuo correspondiente, evidenciando la mayor cantidad de ruido removido y pequeñas porciones adicionales de señal útil.

bandas altas: en la Fig. 5.26 puede verse que, a partir de aproximadamente 14 kHz, el contenido es prácticamente eliminado.

A diferencia de lo observado en los métodos anteriores, el residuo apenas contiene componentes de la voz; esta se conserva casi completamente en la señal procesada y solo puede detectarse mínimamente alrededor del segundo 15. Este comportamiento sugiere que el modelo tiende a preservar con mayor fidelidad las fuentes tonales y armónicas, especialmente la voz, incluso bajo una reducción fuerte del ruido.

En cuanto al carácter del ruido restante, el método elimina casi por completo la granularidad descrita anteriormente y produce un soplido más uniforme, perceptualmente más cercano a un ruido blanco suave. Esta "desgranularización" aporta un fondo más limpio, pero también introduce una cualidad algo más artificial. En términos subjetivos, esto puede percibirse como una ventaja o una desventaja: para algunos oyentes el resultado puede sonar más pulido, mientras que para otros la ausencia de granularidad hace que el soplido remanente se perciba más expuesto, al no estar enmascarado por la textura original.

Es importante subrayar que dicha granularidad no corresponde a artefactos de tipo ruido musical, sino que forma parte de la textura original del audio. Al suprimirla, el método no introduce ruido musical nuevo, como si puede ocurrir con los métodos de sustracción espectral.

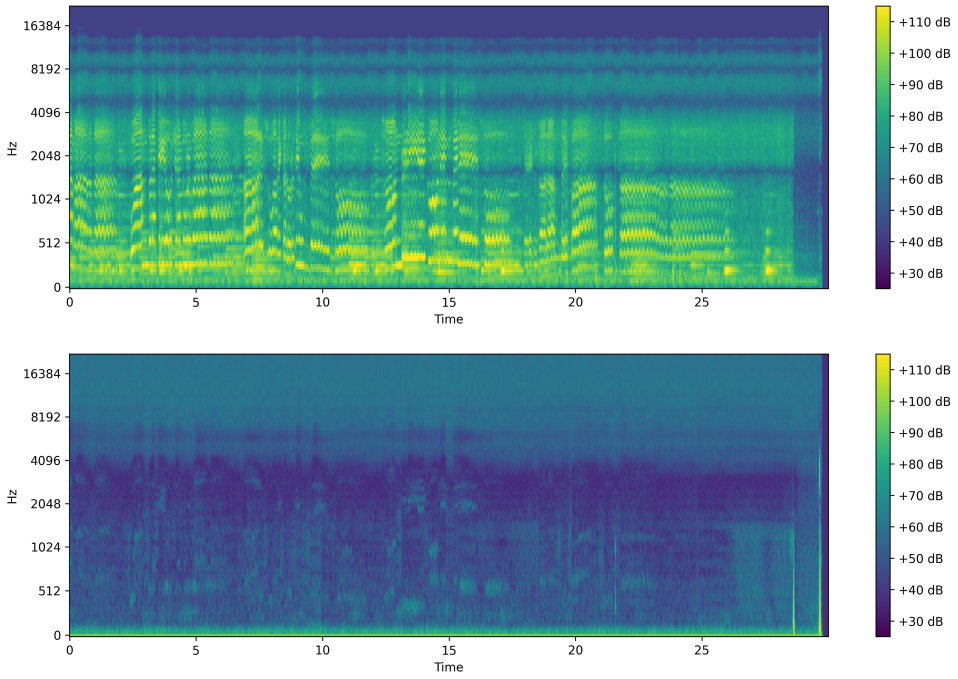

Figura 5.26: Espectrogramas del modelo **DL MagTapeDB**. Arriba: señal procesada, donde se observa la supresión casi total del ruido en las bandas altas. Abajo: residuo generado por el modelo, que muestra la eliminación del ruido granular original y la ausencia de energía vocal significativa.

En términos generales, el modelo realiza un *denoising* eficaz y preserva adecuadamente la voz, pero introduce esta nueva forma de artificialidad que lo distingue de los métodos basados en sustracción espectral. Se trata, nuevamente, de un compromiso perceptual cuya valoración depende del oyente.

Caso 2: "Salite de la esquina" por Rosa Blanca Rodríguez

En 1952, Ayestarán adquiere un grabador de cinta magnética y, a partir de ese momento, sus registros de campo comienzan a realizarse en dicho formato. El fragmento seleccionado forma parte de una serie de canciones infantiles interpretadas por Rosa Blanca Rodríguez, registradas por Ayestarán el 19 de febrero de 1955. Ese día la intérprete grabó al menos seis piezas, entre ellas "Mambrú se fue a la guerra", "La torre en guardia", "Se va, se va la lancha", "En Galicia hay una niña", "En el portal de Belén" y la canción aquí analizada, "Salite de la esquina". No se dispone de información biográfica adicional sobre la cantante fuera de los propios metadatos del archivo.

A diferencia del caso anterior, cuyo origen se remontaba a una grabación en disco instantáneo posteriormente copiada en múltiples soportes, este ejemplo proviene del proyecto *MagTapeDB* [34], descrito en la Sección 4.1.5, una base de datos que reúne digitalizaciones directas de las cintas originales del archivo musicológico de Lauro Ayestarán, evitando así las degradaciones acumuladas observadas en el

caso de estudio anterior.

De esta forma, el ruido presente en este fragmento corresponde a las degradaciones propias de la cinta magnética que se describen en el Anexo B.2. En el espectrograma del audio original, Figura 5.27, se observa, en comparación con el caso anterior, un registro claramente más limpio, lo cual es coherente con el origen de los materiales. El contraste visual es nítido: el fondo aparece mayormente en tonos violetas de baja intensidad, mientras que la voz de la intérprete se distingue con claridad en colores verdes y amarillos.

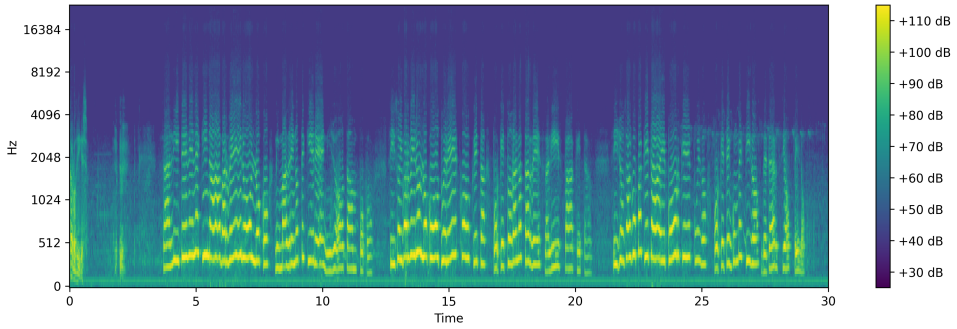

Figura 5.27: Espectrograma del audio original de Rosa Blanca Rodríguez. Se observa un fondo de baja energía, así como la presencia definida de la voz en las bandas medias. En las frecuencias más bajas aparece una línea horizontal persistente, correspondiente a un ruido grave, y en los pasajes de mayor intensidad vocal puede advertirse un leve incremento de energía de banda ancha que rodea los picos de la señal.

A muy bajas frecuencias se aprecia una línea horizontal, correspondiente a un ruido grave, perceptible también en la escucha. Esto podría corresponderse con ruido llamado *hum* descrito en el Anexo B.2, caracterizado por un zumbido estable de baja frecuencia típico de interferencias eléctricas en equipos analógicos. Además, en los pasajes de mayor intensidad vocal se percibe un leve ruido de banda ancha, similar a un ruido blanco, que envuelve los picos de energía de la señal. Este ruido es sutil y requiere cierta atención para advertirse.

En la Figura 5.28 se observa el resultado obtenido mediante *SS Clásico*. Por un lado, el espectrograma superior corresponde a la señal procesada, donde la atenuación del ruido es evidente: la línea horizontal de baja frecuencia desaparece casi por completo y el fondo adquiere una textura significativamente más limpia.

Por otro lado, el espectrograma inferior muestra el residuo, en el que se distingue con claridad la franja horizontal previamente identificada, así como la energía sustraída en bandas más altas, particularmente por debajo de 4 kHz. También aparece una pequeña cantidad de energía asociada a la voz. La escucha confirma estas observaciones: el ruido grave deja de percibirse y la voz pierde la envolvente ruidosa que la acompañaba en el registro original, resultando más clara y definida.

En la escucha aparece también una leve presencia de ruido musical, pero en este caso localizada en las frecuencias bajas. Esto ocurre porque el método realizó la mayor parte de la sustracción por debajo de 4 kHz, que es donde se concentraba el ruido original. A diferencia de lo observado en casos anteriores, donde este

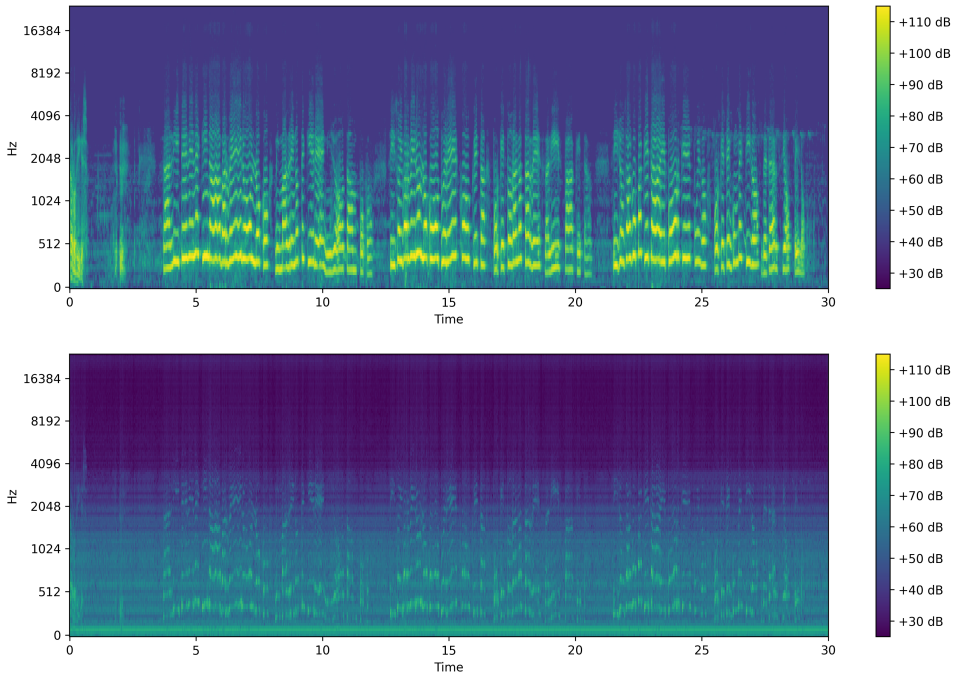

Figura 5.28: Espectrogramas del resultado obtenido mediante **SS Clásico** aplicado al audio de Rosa Blanca Rodríguez. Arriba: señal procesada, donde se observa la reducción del ruido grave y del leve ruido de banda ancha que acompañaba los pasajes más intensos de la voz. Abajo: espectrograma del residuo, que muestra las componentes ruidosas eliminadas y una porción de señal útil retirada.

artefacto surgía en las bandas altas, aquí los puntos aislados aparecen en la zona donde efectivamente se aplicó la atenuación.

En este caso particular, el nivel de ruido presente en la señal original es relativamente bajo, lo que hace que el método clásico ya capture con suficiente precisión las regiones ruidosas. Al tratarse además de un fragmento compuesto únicamente por voz, el comportamiento de *SS Denoisify* no difiere de manera significativa del de *SS Clásico*, incluso al modificar sus hiperparámetros. El resultado obtenido es, en la práctica, muy similar en términos tanto espectrales como perceptuales, por lo que un análisis detallado de esta variante no aportaría elementos nuevos en este contexto.

En el caso del modelo *DL MagTapeDB*, se puede observar en la Figura 5.29 que la señal procesada presenta un espectrograma visualmente distinto al observado en los métodos de sustracción espectral: Se ve un fondo de tonalidad diferente (más cercano al azul que al violeta del espectrograma original), aun cuando no se introduce ruido adicional. Más allá de esa variación de color, el resultado auditivo es notablemente limpio.

Ademas, el modelo opera internamente a 44,1 kHz, por lo que el audio original —registrado a 48 kHz— es remuestreado durante el proceso de inferencia. Este remuestreo no introduce artefactos audibles.

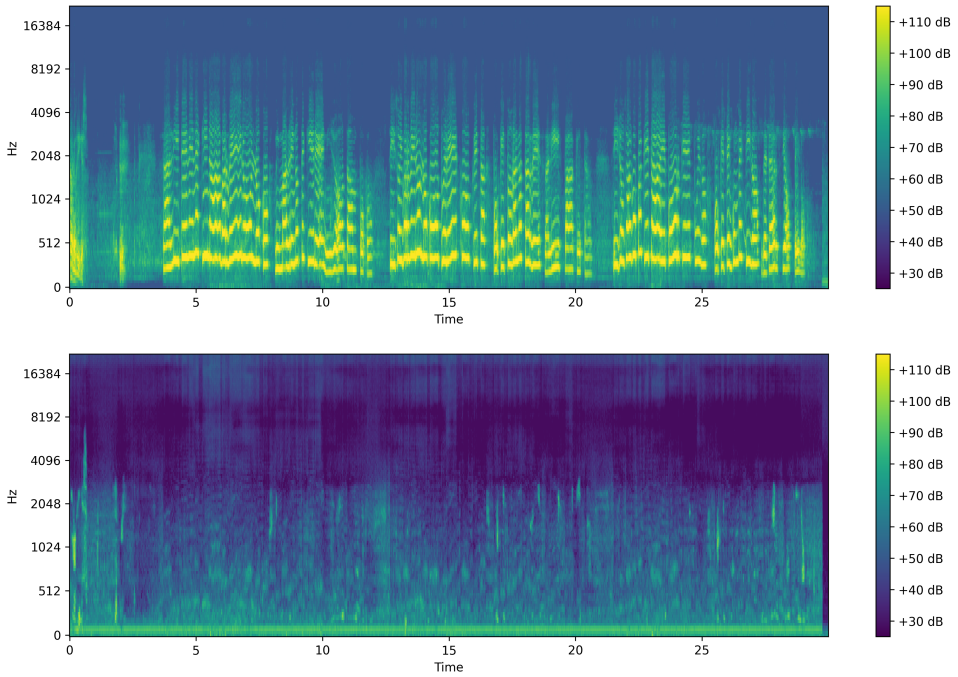

Figura 5.29: Espectrogramas del resultado obtenido mediante **DL MagTapeDB** aplicado al audio de Rosa Blanca Rodríguez. Arriba: señal procesada, donde el fondo se ve ligeramente más azulado, junto con la eliminación del ruido grave presente en la señal original. Abajo: espectrograma del residuo, en el que se distingue claramente la franja de bajas frecuencias y se observan pequeñas componentes por debajo de 4 kHz, junto con una traza muy tenue de la voz.

El modelo elimina con eficacia el ruido grave identificado anteriormente, algo que se aprecia tanto en el espectrograma de salida como en el residuo, donde esa franja de bajas frecuencias aparece completamente aislada. A diferencia de los métodos de sustracción espectral, no se observan artefactos perceptibles: el audio no presenta ruido musical ni irregularidades en las bandas altas. La voz se escucha con claridad y sin la envolvente ruidosa presente en la señal de entrada. En el residuo puede percibirse una traza muy tenue de la voz, aunque considerablemente más atenuada que en el residuo de la sustracción espectral, lo cual indica que el modelo es mejor preservando la señal útil. Desde una perspectiva subjetiva, el resultado constituye una mejora perceptual más marcada que en los métodos basados en sustracción espectral.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 6

Conclusiones

En este capítulo se presentan las conclusiones generales del trabajo, integrando los principales resultados obtenidos y reflexionando sobre su alcance. Para situar adecuadamente estas conclusiones, es pertinente recordar el objetivo central que motivó la investigación: desarrollar, implementar y comparar dos enfoques complementarios para la reducción de ruido en la restauración de grabaciones musicales, combinando técnicas clásicas de procesamiento de señales con modelos basados en aprendizaje profundo.

La motivación que dio origen a este estudio se fundamenta en la necesidad de disponer de herramientas abiertas, comprensibles y reproducibles que permitan mitigar el ruido presente en registros históricos —en particular, en grabaciones en cinta magnética— sin comprometer la información sonora relevante. Este propósito se encuentra directamente vinculado con la preservación del acervo patrimonial asociado a los registros del musicólogo uruguayo Lauro Ayestarán, cuyo valor cultural y documental resalta la importancia de desarrollar metodologías confiables de restauración sonora.

El capítulo se organiza de la siguiente manera: primero, se presentan conjuntamente los principales resultados obtenidos y las limitaciones identificadas tanto para las técnicas de sustracción espectral como para los enfoques basados en aprendizaje profundo. A continuación, se discuten las implicaciones prácticas derivadas de estos hallazgos. Finalmente, se exponen diversas líneas de trabajo futuro que emergen del análisis realizado y que buscan consolidar y ampliar las contribuciones desarrolladas en este estudio.

6.1. Resultados y limitaciones halladas

A partir del análisis conjunto de las métricas objetivas, los tiempos de procesamiento y la escucha crítica realizada, fue posible extraer varias conclusiones generales sobre el comportamiento de los métodos implementados y estudiados.

En primer lugar, los resultados obtenidos muestran que las técnicas clásicas de procesamiento de señales continúan siendo altamente competitivas para la reducción de ruido en grabaciones musicales. En particular, alcanzan los valores

promedio más elevados en el desempeño general según ΔPAQM (1,963 y 1,856), y aunque presentan valores menores en ΔPEAQ (0,392 y 0,202), exponen de las desviaciones estándar más reducidas (0,439 y 0,489). Tanto *SS Clásico* como la implementación *SS Denoisify* logran un equilibrio robusto entre mejora perceptual, estabilidad y costo computacional, manteniendo un desempeño consistente frente a variaciones en la relación señal–ruido y en el tipo de contenido evaluado.

Por otro lado, en los modelos de aprendizaje profundo, cuando el tipo de ruido y el contenido musical de la señal coinciden con los utilizados durante el entrenamiento —como ocurre en el modelo entrenado con *MagTapeDB*— el rendimiento perceptual resulta competitivo e incluso superior en ciertos conjuntos. Por ejemplo, en la categoría *Muchas Fuentes*, el modelo DL MagTapeDB alcanza el mejor valor de $\Delta PAQM$ (2,393) y, como se mencionó anteriormente, dicho conjunto presenta características que se asemejan considerablemente al contenido de la base de datos utilizada para entrenar, *MusicNet*.

No obstante, cuando el modelo se enfrenta a ruidos o contenidos musicales cuyas características difieren de aquellas presentes durante su entrenamiento, su desempeño se deteriora de manera notable, evidenciando una capacidad de generalización limitada frente a condiciones acústicas no contempladas. En la práctica, esto implica la necesidad de recopilar y curar bases de datos especializadas, además de contar con tiempos de entrenamiento prolongados y *hardware* dedicado —generalmente mediante GPU—. Estas exigencias contrastan con la simplicidad y bajo costo computacional de la sustracción espectral.

Además, los tiempos de procesamiento obtenidos evidencian diferencias significativas entre ambas familias de métodos. Mientras que las técnicas clásicas alcanzan tiempos de ejecución reducidos en CPU (0,592 s y 14,774 s), los modelos de aprendizaje profundo requieren intervalos de inferencia considerablemente mayores (aproximadamente 38 s). Esta disparidad condiciona su implementación en sistemas donde el procesamiento en tiempo real constituye un requisito crítico. Sin embargo, los tiempos de inferencia pueden reducirse de forma sustancial cuando se dispone de aceleración mediante GPU.

Por otra parte, la escucha crítica reveló que, si bien tanto la sustracción espectral como el modelo de aprendizaje profundo logran una reducción de ruido perceptible y, en general, satisfactoria, cada técnica introduce patrones de distorsión característicos. La ausencia de una solución universalmente óptima implica que la elección del método debe adecuarse a las particularidades del material y a criterios subjetivos acerca de qué aspectos de la señal se desea priorizar. Vale la pena aclarar que las distorsiones observadas no invalidan la utilidad de estas técnicas, sino que establecen los límites operativos dentro de los cuales cada enfoque resulta más adecuado.

En los modelos de aprendizaje profundo, las distorsiones percibidas resultaron en muchos casos impredecibles, manifestándose como artefactos tonales agudos, atenuación excesiva de componentes graves, eliminación de transitorios o pérdida de detalles en las altas frecuencias. Este comportamiento, altamente dependiente del tipo de señal procesada, carece de un patrón claro y estable, lo que dificulta anticipar el impacto perceptual de la restauración en distintos escenarios.

Por el contrario, las distorsiones asociadas a la sustracción espectral —principalmente el ruido musical—, si bien no son despreciables y pueden afectar la calidad del audio restaurado, presentan un comportamiento mucho más predecible. Su aparición y gravedad pueden anticiparse a partir de las características de la señal de entrada.

Es importante destacar que la severidad de estas distorsiones puede mitigarse mediante la selección adecuada de los parámetros de la implementación, especialmente en el caso de *SS Denoisify*. Esta capacidad de control ofrece una flexibilidad que contrasta con la rigidez de los modelos de aprendizaje profundo, los cuales, una vez entrenados, no permiten modificar el comportamiento de la restauración en caso de que el resultado presente artefactos indeseados. En consecuencia, la sustracción espectral no solo ofrece un comportamiento más estable, sino también un margen de ajuste que facilita su adaptación a distintos tipos de contenido.

En resumen, los resultados obtenidos reflejan un compromiso claro entre flexibilidad, estabilidad y desempeño. Cuando se requiere un método adaptable, con parámetros ajustables y bajo costo computacional, la sustracción espectral es una opción adecuada, ofreciendo un comportamiento estable y consistente. En contextos donde se dispone de bases de datos suficientemente representativas y del hardware necesario para el entrenamiento —preferentemente con aceleración mediante GPU—, los modelos de aprendizaje profundo presentan un potencial considerable. No obstante, la obtención de estos datos supone un desafío significativo, y el desempeño resultante tiende a mostrar una mayor variabilidad e imprevisibilidad. Aun así, con un entrenamiento adecuado, estos modelos pueden alcanzar niveles de restauración superiores, lo que los posiciona como una alternativa prometedora siempre que se satisfagan sus requisitos fundamentales.

6.2. Implicaciones prácticas

El desarrollo realizado en este trabajo no se limita al análisis comparativo entre métodos, sino que también ofrece recursos prácticos orientados a facilitar la comprensión, la experimentación y la reutilización de las herramientas implementadas. En primer lugar, se creó el repositorio *SS Denoisify*, disponible públicamente en GitHub [48]. Allí se incluye la implementación completa del algoritmo de sustracción espectral desarrollada a lo largo del trabajo, organizada en distintos módulos y *scripts* de `Python`. Además, el repositorio incorpora un *Jupyter Notebook* diseñado con fines didácticos, en el cual se ilustra paso a paso el funcionamiento del método y se muestra el proceso de restauración de una señal ruidosa de ejemplo.

Además, con el objetivo de acompañar el informe y facilitar la exploración de los resultados obtenidos, se desarrolló una página web interactiva [44] donde es posible visualizar las señales originales utilizadas en el análisis, junto con sus correspondientes versiones restauradas mediante cada uno de los métodos evaluados. El sitio incluye también espectrogramas comparativos y ejemplos auditivos que permiten apreciar de forma directa las diferencias entre las técnicas.

Por último, este trabajo destaca la vigencia y el valor del procesamiento tradicional de señales en un contexto dominado por el aprendizaje automático. Si bien

las técnicas basadas en aprendizaje profundo constituyen hoy en día el principal foco de investigación y aplicación en tareas de restauración de audio, los resultados obtenidos muestran que los métodos clásicos siguen siendo herramientas relevantes. Su eficiencia, estabilidad y ausencia de requisitos de entrenamiento los convierten en alternativas especialmente valiosas en escenarios donde la obtención de datos adecuados es difícil o directamente inviable. En este sentido, este proyecto contribuye a reivindicar el rol del procesamiento tradicional como un enfoque plenamente vigente, capaz de ofrecer soluciones sólidas en aplicaciones donde los modelos de aprendizaje profundo no resultan prácticos.

6.3. Líneas futuras de trabajo

A partir de los resultados obtenidos, las limitaciones identificadas y las implicaciones prácticas del estudio, esta sección presenta diversas propuestas de líneas de investigación futura orientadas a profundizar, complementar y mejorar los desarrollos realizados en el presente trabajo.

6.3.1. Detección de inactividad de la señal

En primer lugar, tal como exhiben los resultados obtenidos en la búsqueda de hiperparámetros del algoritmo de detección de inactividad, un valor de *precision* de 62,03 % indica que una proporción considerable de *frames* con actividad relevante de la señal fue clasificada erróneamente como inactiva. Como se pudo observar, estas detecciones incorrectas afectan negativamente el cálculo del perfil de ruido empleado en los procesos de sustracción espectral. Si bien el análisis permitió identificar que este fenómeno puede deberse, en parte, al enmascaramiento de la señal por el ruido, es fundamental mejorar la *precision*.

En este sentido, una línea de trabajo a futuro consiste en explorar estrategias alternativas para la detección de inactividad que no se limiten exclusivamente al dominio musical o del audio, con el fin de identificar e implementar un algoritmo potencialmente más robusto que el presentado en la Sección 2.4.

Por otra parte, dado el creciente desarrollo de los modelos de aprendizaje profundo, también podría considerarse el diseño o adopción de una arquitectura neuronal capaz de identificar automáticamente los segmentos inactivos de una señal. Sin embargo, este enfoque requeriría un proceso exhaustivo de etiquetado de datos para definir con precisión los intervalos de silencio y diseñar una función de pérdida adecuada, lo que representa un desafío considerable debido a la gran cantidad de ejemplos necesarios para lograr un entrenamiento y un aprendizaje efectivo.

6.3.2. Combinación de ambas técnicas

Como se ha señalado a lo largo del análisis, ambas familias de métodos presentan limitaciones propias. La sustracción espectral, si bien efectiva y flexible, tiende a generar ruido musical con características bien definidas —descritas en

la Subsección 2.5.1—, mientras que los modelos de aprendizaje profundo dependen fuertemente de la disponibilidad de bases de datos que representen fielmente el dominio de las señales que se desean restaurar, lo cual constituye un desafío complejo.

En este escenario, una línea de trabajo futura especialmente prometedora consiste en explorar enfoques híbridos que integren ambas metodologías. Una posibilidad es aplicar primero una sustracción espectral para reducir el ruido de fondo —aceptando la aparición de ruido musical— y, en una segunda etapa, utilizar un modelo de aprendizaje profundo específicamente entrenado para suprimir este artefacto. Para ello podría emplearse una base de datos generada artificialmente, donde las señales limpias sean degradadas únicamente mediante la introducción controlada de ruido musical.

Bajo la hipótesis de que el ruido musical surge principalmente del propio mecanismo de sustracción espectral y no del tipo de ruido original (cinta, gramófono u otras fuentes), este enfoque permitiría entrenar un único modelo capaz de eliminar sistemáticamente dicho artefacto en una amplia variedad de situaciones. De este modo, el sistema resultante combinaría la capacidad generalizada de reducción de ruido de la sustracción espectral con la potencia de los modelos neuronales para refinar el resultado final, sin requerir grandes bases de datos específicas para cada escenario de degradación real.

6.3.3. Desarrollo de bases de datos para el entrenamiento

En continuidad con la propuesta anterior, futuras líneas de trabajo deberían orientarse a la construcción o ampliación de bases de datos de ruido que abarquen distintos soportes y contextos históricos, incluyendo no solo la captura del ruido residual propio de cada medio, sino también su variabilidad asociada al envejecimiento, las condiciones ambientales y los procesos de digitalización.

Una estrategia complementaria podría consistir en ampliar las bases existentes mediante técnicas de *data augmentation*, generando ejemplos sintéticos que emulen degradaciones típicas del audio analógico. La aplicación controlada de estas transformaciones permitiría diversificar los escenarios de entrenamiento y aumentar la capacidad de generalización de los modelos.

6.3.4. Dinámica del aprendizaje del modelo

Del análisis de las curvas de aprendizaje surge un aspecto importante: el desempeño de un modelo de aprendizaje profundo no puede evaluarse únicamente mediante métricas como el MAE. Si bien esta medida cuantifica la discrepancia promedio entre el espectrograma estimado y su referencia limpia, no refleja necesariamente la presencia de artefactos perceptuales ni la calidad subjetiva del audio resultante. En consecuencia, un MAE reducido no implica, por sí mismo, una restauración auditivamente satisfactoria.

Tal como se expuso en capítulos previos, este trabajo incorporó métricas perceptuales como PEAQ y PAQM para complementar esas limitaciones. No obstan-

te, dichas métricas pueden ofrecer valoraciones diferentes sobre qué constituye una mejora perceptual. Tal como se observó en el análisis objetivo, cada una se basa en criterios distintos y, en consecuencia, no necesariamente evalúan la calidad sonora de la misma manera.

En este contexto, una línea futura de investigación consiste en desarrollar esquemas de entrenamiento donde la función de pérdida integre directamente criterios perceptuales más estrechamente vinculados con la escucha humana. Esto permitiría orientar el proceso de aprendizaje hacia mejoras cuantitativas que se correlacionen de manera más consistente con la experiencia auditiva real. Sin embargo, es importante señalar que la restauración de audio continúa siendo una tarea intrínsecamente subjetiva: distintos oyentes —así como distintas métricas— pueden priorizar atributos diferentes del sonido. En consecuencia, la determinación de un "mejor" resultado constituye un desafío complejo y, en muchos casos, dependiente del criterio adoptado.

Apéndice A

Análisis de las métricas para la detección de inactividad

Si bien existe una amplia bibliografía sobre detección automática de inactividad, la mayor parte de los desarrollos se enfocan en señales de voz, donde la detección de actividad vocal es una herramienta esencial en aplicaciones de telecomunicaciones, codificación de voz y asistentes virtuales [49–56]. No obstante, estos métodos están diseñados para las características propias del habla y no se adaptan directamente a señales musicales, lo que motiva la necesidad de una solución específica para el contexto abordado en este trabajo.

A continuación, se analizan y evalúan algunas de las métricas y características empleadas en dichos trabajos, con el fin de identificar cuáles de ellas pueden ser adaptadas e integradas en el módulo de detección de inactividad desarrollado para este proyecto.

A.1. Técnicas implementadas para VAD

Los algoritmos de detección de actividad de voz han evolucionado notablemente desde sus primeras propuestas. Inicialmente se emplearon enfoques heurísticos basados en características simples de la señal, lo que permitió implementaciones eficientes en tiempo real. Entre los trabajos pioneros destaca Atal y Rabiner (1976) [52], quienes propusieron un método de clasificación de segmentos de voz mediante parámetros acústicos como la energía, los cruces por cero o coeficientes LPC, logrando una segmentación eficaz incluso en intervalos cortos. Posteriormente, Tucker (1992) [53] introdujo un algoritmo basado en la estimación de periodicidad, robusto en condiciones de bajo SNR. Otros enfoques relevantes incorporaron análisis cepstral para discriminación entre habla y ruido [54], así como medidas de entropía espectral para robustez en entornos ruidosos [55, 56].

Además, se desarrollaron técnicas basadas en energía y umbrales dinámicos [57–59], que ajustan la detección en función de las variaciones del entorno acústico. Estas estrategias, aunque simples, continúan siendo útiles por su bajo costo computacional.

A.2. Análisis

Si bien la detección de actividad en señales de voz humana es un campo ampliamente estudiado, las características propias de este tipo de señales no siempre se trasladan directamente a otros dominios, como el musical. Las grabaciones de voz presentan rasgos distintivos, entre los que se destacan: la presencia habitual de silencios prolongados, una concentración energética predominante en las bajas frecuencias, una densidad espectral menos distribuida, variaciones de volumen generalmente suaves y graduales, y la emisión desde una única fuente sonora o, en su defecto, desde un número muy limitado de fuentes simultáneas.

Las señales musicales no necesariamente comparten estas particularidades. Por ejemplo, en una orquesta se encuentran una gran variedad de instrumentos con diferentes timbres y rangos sonoros, lo que hace que este tipo de señales sea considerablemente más complejo.

Por otro lado, existen diferencias fundamentales en cuanto al propósito del algoritmo de detección según el contexto de aplicación. Por ejemplo, en las telecomunicaciones, el propósito principal es identificar los *frames* que contienen actividad vocal con el fin de evitar la transmisión innecesaria de datos durante los períodos de silencio o ruido de fondo. Esto permite reducir significativamente la cantidad de información enviada, optimizando el uso del ancho de banda sin comprometer la inteligibilidad del mensaje. En ese escenario, la pérdida ocasional de algunos *frames* de audio no suele ser crítica, siempre que no sea perceptible para el oyente.

En cambio, en el enfoque adoptado en este trabajo, el interés no radica en preservar la señal útil, sino en obtener muestras representativas del ruido de fondo. Por ello, no es prioritario detectar todos los *frames* de silencio, sino garantizar que los *frames* seleccionados correspondan efectivamente a segmentos sin actividad útil.

Otra diferencia importante radica en la hipótesis sobre la estacionariedad del ruido. En los algoritmos descritos en [57–59], los umbrales utilizados para la detección de segmentos activos o pasivos se actualizan dinámicamente, bajo la suposición de que el ruido es aditivo y localmente estacionario. Es decir, se asume que durante los breves períodos en los que una persona está hablando, las características estadísticas del ruido de fondo permanecen aproximadamente constantes.

Sin embargo, esta suposición no se traslada fácilmente al contexto de las grabaciones musicológicas. A diferencia del habla, que está naturalmente segmentada por pausas y silencios, la música suele presentar fragmentos extensos sin interrupciones marcadas, lo que dificulta identificar regiones sin actividad útil. En consecuencia, la hipótesis de ruido localmente estacionario resulta menos adecuada, y debe ser reemplazada por una condición más estricta: que el ruido sea aproximadamente estacionario a lo largo de toda la señal. Esta restricción es considerablemente más exigente y plantea nuevos desafíos para la detección y estimación del perfil de ruido.

A partir de las observaciones realizadas, se decidió analizar las métricas utilizadas en los algoritmos previamente mencionados para la detección de inactividad,

con el objetivo de adaptarlas a las características específicas del presente caso. Las métricas evaluadas fueron las siguientes:

- Energía en tiempo corto.
- Periodicidad.
- Métricas basadas en la Autocorrelación entre Muestras.
- Métricas basadas en la Entropia espectral.
- Métricas basadas en Cepstrum.
- Tasa de cruces por cero en tiempo corto.

A.2.1. Energía en tiempo corto

En el contexto de señales musicales contaminadas con ruido, esta métrica resulta especialmente útil para detectar períodos de inactividad siempre que la SNR sea razonable. Bajo estas condiciones, la señal musical suele dominar sobre el ruido durante los tramos activos, generando una energía significativamente mayor que en los momentos de silencio. De este modo, los intervalos donde solo persiste el ruido tienden a presentar niveles de energía más bajos y estables, lo que permite separarlos mediante umbrales adecuados.

La energía en tiempo corto tiene la ventaja de ser una métrica sencilla de calcular y relativamente robusta frente a variaciones armónicas o instrumentales propias de la música, siempre que el ruido no sea excesivamente intrusivo. Por estas razones, ha sido empleada de forma recurrente en distintos trabajos de detección de actividad [52, 57–59].

A.2.2. Periodicidad

La aplicabilidad de los estimadores de periodicidad, como el basado en mínimos cuadrados propuesto en [53], resulta limitada en el contexto de grabaciones musicales polifónicas. Cuando múltiples instrumentos suenan simultáneamente, cada uno con sus propias características espectrales y temporales, la señal resultante carece de una única periodicidad, convirtiéndose en una superposición densa de distintas componentes. Esto dificulta la identificación de un patrón periódico principal y puede provocar estimaciones inestables o erróneas.

También, en grabaciones que contienen instrumentos de percusión, los cuales pueden implicar la presencia de transitorios abruptos y variaciones rápidas, la capacidad del modelo para representar adecuadamente la estructura periódica de la señal se ve aún más comprometida. Por estas razones, el uso de la periodicidad como criterio para la detección de inactividad en grabaciones musicológicas resulta poco confiable.

A.2.3. Métricas basadas en la autocorrelación entre muestras

Las métricas basadas en la autocorrelación entre muestras de la señal, tales como *Maximum Autocorrelation Peak*, *Autocorrelation Peak Count* o *Windowed Autocorrelation Lag Energy*, abordadas en [52, 60–62], pueden resultar útiles en el análisis de señales de voz, así como en grabaciones musicales con pocas fuentes sonoras.

Sin embargo, al igual que ocurre con las métricas basadas en la periodicidad, su desempeño se ve limitado en grabaciones con múltiples instrumentos sonando simultáneamente, ya que las señales generadas por distintas fuentes no necesariamente presentan una correlación temporal entre sí. Las grabaciones musicales suelen ser altamente dinámicas y estructuralmente variadas, lo que complica aún más la detección de correlaciones consistentes entre muestras, reduciendo la confiabilidad de estas métricas para identificar segmentos con actividad o con solo ruido de fondo en el contexto del presente trabajo.

A.2.4. Métricas basadas en la entropía espectral

La entropía espectral se basa en interpretar la distribución de energía en frecuencia de una señal como una distribución de probabilidad, sobre la cual se calcula la entropía de *C. E. Shannon* [63]. De este modo, esta métrica describe cuán dispersa o concentrada está la energía en el dominio espectral. Una alta entropía espectral indica que la energía está distribuida de manera relativamente uniforme a lo largo de las frecuencias, lo que sugiere una señal con amplio contenido frecuencial y sin componentes dominantes. En cambio, una baja entropía espectral refleja que la energía se encuentra concentrada en unas pocas frecuencias, lo que es característico de señales tonales o estructuradas.

Esta métrica puede resultar útil en escenarios donde el ruido de fondo es aproximadamente blanco, ya que en tales casos la energía del ruido está distribuida de manera uniforme en el espectro, generando una entropía espectral alta y fácilmente distinguible de señales estructuradas.

Sin embargo, en el presente trabajo, el ruido no es necesariamente blanco ni gaussiano, y puede presentar una distribución espectral con picos de energía en ciertas bandas. Como consecuencia, la entropía espectral del ruido puede no diferir sustancialmente de la de la señal de interés, lo que reduce la capacidad discriminatoria de esta métrica para detectar actividad frente a ruido de fondo.

En la Figura A.1 se presentan los espectrogramas de tres señales de ruido de cinta, extraídas del trabajo [4]. Dicho estudio considera el ruido generado por distintos dispositivos de grabación, entre los que se incluyen el *Revox A77*, el *Uher 4000 Report L* y el *Technics TR-575*, cuyas características espectrales pueden observarse en la figura mencionada. Como se aprecia, los espectros no son uniformes ni planos, lo que los distingue de un ruido blanco o gaussiano. En consecuencia, métricas como la entropía espectral no resultan completamente efectivas para detectar actividad en señales contaminadas con estos tipos de ruido.

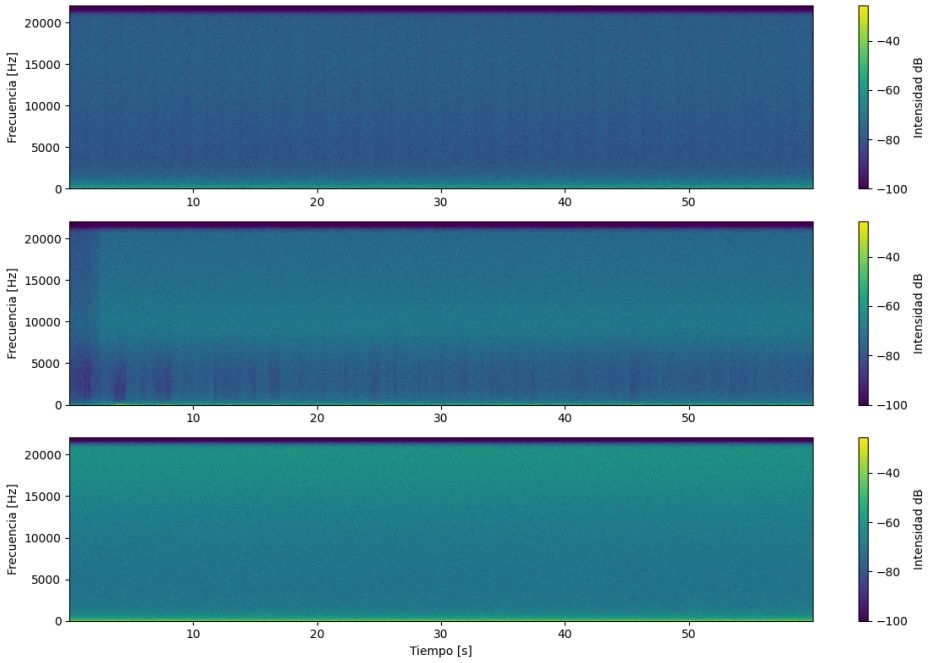

Figura A.1: Espectrogramas de señales de ruido de cinta normalizadas, obtenidas a partir de tres grabadores analógicos (*Revox A77*, *Uher 4000 Report L* y *Technics TR-575*) [4]. Se observa que las características espectrales difieren de las de un ruido blanco o gaussiano ideal, presentando distribuciones no uniformes. Esta particularidad dificulta la aplicación de métricas clásicas como la entropía espectral para la detección de actividad en señales contaminadas con este tipo de ruido.

A.2.5. Métricas basadas en Cepstrum

El *cepstrum* es una representación de señales obtenida al aplicar la transformada de Fourier al logaritmo del espectro de magnitud de una señal. Esta transformación permite analizar estructuras periódicas en el dominio de la frecuencia, como la detección de armónicos en señales de audio o la separación entre la envolvente espectral y la señal de excitación en sistemas acústicos. Una de sus propiedades fundamentales es que convierte convoluciones en el dominio temporal en sumas en el dominio de las *quefrencies*. Esto se debe a que, en el dominio de Fourier, una convolución temporal se traduce en un producto espectral, y al aplicar el logaritmo, dicho producto se transforma en una suma, lo cual facilita la separación de componentes como la envolvente y la estructura armónica.

Cuando la señal proviene de múltiples fuentes la representación cepstral se vuelve más compleja debido a la superposición de múltiples estructuras periódicas. Esta superposición genera varios picos en diferentes valores de *quefrency*, dificultando la identificación de una periodicidad dominante, al igual que en el caso de las métricas basadas en la autocorrelación entre muestras y en la periodicidad. Como consecuencia, los coeficientes cepstrales tienden a dispersarse, lo que complica su interpretación y reduce la robustez de estas métricas para el caso de uso considerado en este trabajo.

A.2.6. Taza de cruces por cero en tiempo corto

La tasa de cruces por cero (*Zero Crossing Rate*, ZCR) es una métrica que cuantifica cuántas veces una señal cambia de signo en un intervalo de tiempo determinado. En el contexto del procesamiento de señales de audio, se calcula usualmente en ventanas de tiempo cortas y representa la cantidad de veces que la señal cruza el eje horizontal (pasa de positiva a negativa o viceversa). Una ZCR alta indica una señal con componentes de frecuencia elevada o con variaciones rápidas, como ocurre en el ruido blanco, mientras que una ZCR baja sugiere la presencia de componentes de baja frecuencia o una señal más suave. Por su simplicidad y bajo costo computacional, esta métrica ha sido utilizada para la detección de actividad en señales de audio como en [52].

Cuando el ruido presente en la señal tiene media aproximadamente nula y componentes de alta frecuencia, la ZCR resulta especialmente útil para detectar actividad en señales musicales. Por ejemplo, al combinar la ZCR con la energía de la señal, se pueden identificar como ruido los segmentos que presentan energía casi nula pero una ZCR alta. Esta metodología es efectiva siempre que la energía del ruido sea significativamente menor que la de la señal (es decir, que la SNR, sea suficientemente alta) y que el ruido tenga una media cercana a cero. Sin embargo, es importante tener precaución, ya que ciertos fragmentos agudos de la señal, también pueden exhibir una ZCR alta, lo que podría ocasionar errores en la detección si no se manejan adecuadamente.

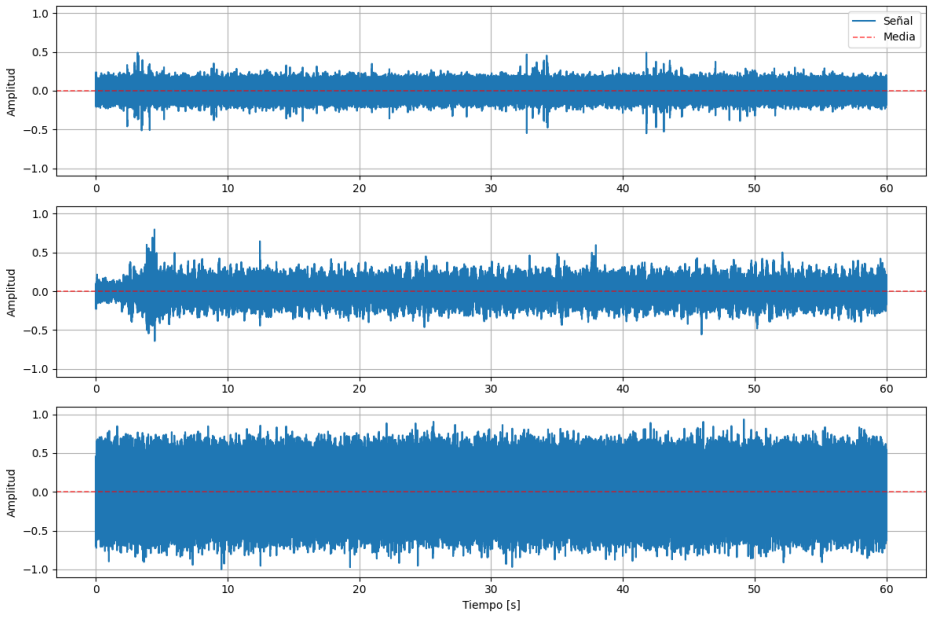

Figura A.2: Señales temporales normalizadas de ruido de cinta, correspondientes a los grabadores analógicos *Revox A77*, *Uher 4000 Report L* y *Technics TR-575*, respectivamente. Se aprecia que las señales presentan un comportamiento aproximadamente estacionario, con componentes de alta frecuencia y valores medios cercanos a cero, en concordancia con las hipótesis asumidas en el análisis.

En la Figura A.2 se presentan las señales temporales y sus valores medios correspondientes a los ruidos de cinta mencionados anteriormente. Junto con la información mostrada en la Figura A.1, se puede observar que estos ruidos son aproximadamente estacionarios, que contienen componentes de alta frecuencia y que tienen una media cercana a cero, cumpliendo así con las hipótesis planteadas previamente.

Esta página ha sido intencionalmente dejada en blanco.

Apéndice B

Descripción de los tipos de ruido en soportes históricos

Este anexo presenta una caracterización de los ruidos típicos presentes en dos soportes analógicos de relevancia histórica: los discos de 78 RPM (gramófono) y las cintas magnéticas. El objetivo es documentar las degradaciones propias de cada medio y establecer el marco acústico en el que operan los métodos de restauración estudiados en esta tesis. Si bien ambos soportes comparten problemas generales relacionados con el envejecimiento y las limitaciones tecnológicas de la época, cada uno introduce artefactos específicos derivados de su naturaleza física y su mecanismo de reproducción.

B.1. Ruido característico de los discos de gramófono (78 RPM)

Las grabaciones de discos de 78 RPM presentan una variedad de artefactos acústicos propios del soporte y del equipamiento de reproducción. En primer lugar, el *hiss* corresponde a un ruido de banda ancha generado por las etapas analógicas del sistema (preamplificadores, circuitería y ruido térmico), que se manifiesta como un siseo constante con mayor concentración de energía en las altas frecuencias. Otro componente característico es el *rumble*, un zumbido de muy baja frecuencia producido por vibraciones mecánicas del motor, desalineaciones del eje o resonancias estructurales del giradiscos; su energía se concentra típicamente por debajo de 80–100 Hz y resulta especialmente audible en pasajes silenciosos.

Además del ruido continuo, los discos de goma laca suelen presentar artefactos impulsivos. Los *clicks* son chasquidos de muy corta duración originados por rayaduras finas, acumulación de polvo o microfisuras en el surco; su espectro es amplio y con fuerte contenido de alta frecuencia, lo que los hace perceptualmente agudos y bien definidos. Por su parte, los *thumps* son golpes más largos y predominantemente de baja frecuencia, asociados a deformaciones más profundas del surco, daños estructurales o impactos en la cápsula durante la reproducción. Finalmente, las digitalizaciones históricas también incorporan ruido ambiental propio del entorno

de captura —como ventilación, interferencias eléctricas o vibraciones transmitidas al plato— que añade componentes adicionales de fondo con un espectro variable según la fuente.

Estos artefactos conforman un perfil de ruido altamente heterogéneo y dependiente del estado físico del disco. La ausencia de estandarización en los procesos de grabación y reproducción de la época añade aún más variabilidad, lo que convierte al dominio del gramófono en un entorno difícil para la restauración automática.

B.2. Ruido característico en cinta magnética

En contraste con el gramófono, la cinta magnética introduce degradaciones menos asociadas al desgaste del soporte físico superficial y más vinculadas a los principios electromagnéticos del registro analógico [1]. A continuación, se describen en mayor detalle los fenómenos relevantes para la restauración y el análisis desarrollados en este trabajo.

B.2.1. Ruido de banda ancha (*hiss*)

El *hiss* es un ruido de fondo continuo, similar a un "siseo", que aparece de forma natural en las grabaciones en cinta magnética. Proviene del propio material del soporte: con el paso del tiempo, la superficie de la cinta desarrolla pequeñas irregularidades y variaciones aleatorias que generan ruido incluso cuando no hay señal registrada [1]. Este efecto se vuelve más evidente a medida que la cinta envejece o pierde estabilidad magnética [2]. El hiss se concentra sobre todo en las frecuencias altas y suele estar presente en toda la duración de la grabación.

B.2.2. Interferencias eléctricas (*hum*)

El *hum* es un zumbido de baja frecuencia que aparece cuando la grabadora o el reproductor captan interferencias provenientes de la red eléctrica [2]. Suele escucharse en 50 Hz o 60 Hz, junto con sus armónicos, y puede deberse a transformadores, motores del transporte de cinta, fuentes de alimentación desgastadas o problemas de masa. Este ruido resulta especialmente molesto en pasajes silenciosos o con poca dinámica.

B.2.3. Inestabilidades de velocidad (*wow and flutter*)

Las grabaciones en cinta también pueden sufrir variaciones en la velocidad de arrastre, lo que produce fluctuaciones audibles en el tono. El *wow* corresponde a cambios lentos y periódicos en la velocidad, generalmente provocados por problemas mecánicos en el capstan, los rodillos o las guías del transporte [64]. El *flutter*, por su parte, es una variación más rápida e irregular causada por desgaste en el motor, rodillos endurecidos o vibraciones de la estructura [1]. Ambos efectos generan pequeñas oscilaciones del tono que afectan la estabilidad y naturalidad del sonido.

B.2.4. Saturación magnética y distorsión

La saturación ocurre cuando la señal registrada es demasiado intensa y el soporte ya no puede almacenar más magnetización [1]. En ese punto, la cinta deja de responder de forma lineal y aparecen distorsión armónica, recorte y una compresión no deseada del sonido. Este problema suele deberse a una mala calibración durante la grabación o a intentos de aumentar la relación señal–ruido llevando el nivel demasiado cerca del límite físico de la cinta [2].

B.2.5. Caídas de señal (*dropouts*)

Los *dropouts* son pequeñas pérdidas momentáneas de audio que ocurren cuando alguna parte de la capa magnética de la cinta está dañada o debilitada [65]. Pueden deberse a abrasión, suciedad, moho o defectos mecánicos de la superficie. Como la información se almacena en una capa muy fina, cualquier interrupción en esa zona provoca una caída abrupta del nivel registrado. Según su tamaño y duración, los dropouts pueden percibirse como breves "huecos" en el sonido, cambios de timbre o pérdidas instantáneas de alta frecuencia.

B.2.6. Degradación química del aglutinante (*sticky-shed syndrome*)

Otro fenómeno frecuente en cintas antiguas es el *sticky-shed syndrome* o "síndrome de la capa pegajosa". Ocurre cuando el material que mantiene adheridas las partículas magnéticas al soporte comienza a degradarse con el paso del tiempo. La cinta absorbe humedad y el aglutinante se descompone, haciendo que la superficie se ablande y se vuelva pegajosa [64]. Esto provoca fricción excesiva durante la reproducción, acumulación de residuos en los cabezales e incluso, en casos graves, que la cinta no pueda reproducirse sin riesgo de dañarla.

En el sonido, este problema puede percibirse como ruidos intermitentes, pérdida de agudos y pequeñas variaciones de velocidad debido al arrastre irregular. Su tratamiento requiere procedimientos de conservación específicos, como el secado controlado o "horneado" previo a la digitalización, una solución que permite reproducir la cinta de forma temporal pero no detiene su deterioro a largo plazo.

B.2.7. Otros artefactos relevantes

Además de los problemas principales, el deterioro químico de la cinta puede generar otros efectos, como variaciones en el azimut, ruidos provocados por un deslizamiento irregular sobre los cabezales o aumentos de fricción que afectan la velocidad de arrastre [64, 66]. También los propios cabezales pueden degradarse con el uso, lo que provoca pérdidas de altas frecuencias debido a desgaste físico o desmagnetización parcial [1].

Esta página ha sido intencionalmente dejada en blanco.

Referencias

[1] R. L. Hess, "Tape degradation factors and challenges in predicting tape life," tech. rep., Association for Recorded Sound Collections (ARSC), 2011. Estudio sobre la degradación de cintas magnéticas analógicas; "each duplication generates a small degradation in quality".

[2] M. J. Kromhout, ""how much noise is necessary?" a brief history of audio tape noise reduction," tech. rep., University of Amsterdam, 2021. Capítulo donde se revisa la evolución de sistemas de reducción de ruido en cinta analógica y se mencionan los artefactos asociados a dicha tecnología.

[3] Centro de Documentación Musical Lauro Ayestarán, "Centro de documentación musical lauro ayestarán," 2024. Disponible en: `http://www.cdm.gub.uy/`.

[4] I. Irigaray, M. Rocamora, and L. W. P. Biscainho, "Noise reduction in analog tape audio recordings with deep learning models," *Journal of the Audio Engineering Society*, June 2023. Presented at the AES International Conference on Audio Archiving, Preservation & Restoration, Culpeper, VA, USA, June 1–3, 2023.

[5] I. Irigaray and F. Sallés, "Aproximación interdisciplinaria al trabajo con documentos sonoros. estudio de caso: las grabaciones de campo de lauro ayestarán," tech. rep., Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, UDELAR, 2019.

[6] E. Moliner and V. Välimäki, "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, 2022.

[7] E. Moliner and V. Välimäki, "Behm-gan: Bandwidth extension of historical music using generative adversarial networks," *IEEE Transactions on Audio, Speech, and Language Processing*, 2022.

[8] iZotope, Inc., "Rx 11 - the definitive toolkit for audio repair," 2024.

[9] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[10] S. Ogata and T. Shimamura, "Reinforced spectral subtraction method to enhance speech signal," in *Proceedings International Conference on Electrical and Electronic Technology (ICEET)*, vol. 1, pp. 242–245, 2001.

[11] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Computer Science*, vol. 54, pp. 574–584, 2015. Eleventh International Conference on Communication Networks, ICCN 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Bangalore, India.

[12] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

[13] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*. Chichester, UK: John Wiley & Sons, 4 ed., 2008.

[14] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Washington, DC), pp. 208–211, April 1979.

[15] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Orlando, USA), May 2002.

[16] N. Upadhyay and A. Karmakar, "Single channel speech enhancement utilizing iterative processing of multi-band spectral subtraction algorithm," in *2012 2nd International Conference on Power, Control and Embedded Systems*, pp. 1–6, 2012.

[17] M. El-Fattah, M. Dessouky, S. Diab, and F. Abd El-Samie, "Speech enhancement using an adaptive wiener filtering approach," *Progress in Electromagnetics Research M*, vol. 4, pp. 167–184, 01 2008.

[18] S. Li, "Iterative spectral subtraction method for millimeter-wave conducted speech enhancement," *Journal of Biomedical Science and Engineering*, vol. 03, pp. 187–192, 01 2010.

[19] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of iterative weak spectral subtraction via higher-order statistics," in *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 220–225, 2010.

[20] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on

higher order statistics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1770–1779, 2011.

[21] J. Xiaoping, F. Hua, and Y. Tianren, "Single-channel speech enhancement method based on masking properties and minimum statistics," vol. 15, pp. 460 – 463 vol.1, 09 2002.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[23] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[24] X. Serra, *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition.* Phd dissertation, Stanford University, Stanford, 1989.

[25] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[26] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing* (C. Roads, S. Pope, A. Picialli, and G. de Poli, eds.), pp. 91–122, Lisse: Swets & Zeitlinger, 1997.

[27] X. Serra and J. Bonada, "Sound transformations based on the sms high level attributes," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, (Barcelona, Spain), 1998.

[28] Music Technology Group (MTG), "Sms tools - spectral modeling synthesis tools," 2024. GitHub repository. Accessed: 2025-07-30.

[29] X. Amatriain, J. Bonada, Loscos, and X. Serra, "Spectral processing," in *DAFX - Digital Audio Effects* (U. Zölzer, ed.), pp. 373–438, Chichester: John Wiley & Sons, 2002.

[30] D. Fitzgerald, "Harmonic/percussive separation using median filtering," *13th International Conference on Digital Audio Effects (DAFx-10)*, 01 2010.

[31] M. Müller, "Harmonic–percussive separation (hpss)," 2023. FMP Notebooks, AudioLabs Erlangen. Accessed: 2025-07-30.

[32] L. Moysis, L. A. Iliadis, S. P. Sotiroudis, A. D. Boursianis, M. S. Papadopoulou, K.-I. D. Kokkinidis, C. Volos, P. Sarigiannidis, S. Nikolaidis, and S. K. Goudos, "Music deep learning: Deep learning methods for music signal processing—a review of the state-of-the-art," *IEEE Access*, vol. 11, pp. 17031–17052, 2023.

[33] Y. Li, B. Gfeller, M. Tagliasacchi, and D. Roblek, "Learning to denoise historical music," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, (Montréal, Canada), pp. 504–511, Oct. 2020.

[34] I. Irigaray, E. Martinez, D. S. Coeff, and P. B. Luiz W. "Magtapedb: A dataset of historical magnetic tape recordings," in *Proceedings of the IEEE 15th International Conference on Pattern Recognition Systems (ICPRS 2025)*, (Viña del Mar, Chile), IEEE / IAPR, 2025.

[35] J. Thickstun, Z. Harchaoui, and S. M. Kakade, "Learning features of music from scratch," in *Proceedings of the International Conference on Learning Representations (ICLR)*, (Toulon, France), Apr. 2017.

[36] "The great 78 project." `https://great78.archive.org`. Accessed: February 18, 2022.

[37] U. İsik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," *arXiv preprint arXiv:2008.04470*, 2020. Se introduce el uso de *frequency-positional embeddings* para mejorar la capacidad del modelo de desarrollar características dependientes de la frecuencia en las primeras capas.

[38] International Telecommunication Union (ITU), "Recommendation itu-r bs.1387-1: Method for objective measurements of perceived audio quality," 1998. Disponible en: `https://www.itu.int/rec/R-REC-BS.1387`.

[39] HSU-ANT, "Gst-peaq: A perceptual audio quality model implementation," 2024. Disponible en: `https://github.com/HSU-ANT/gstpeaq`.

[40] GitHub Repository, "Implementation of paqm in python using pytorch for vectorized operations." `https://github.com/bvm810/paqm-python.git`, 2024. Disponible en: `https://github.com/bvm810/paqm-python.git`.

[41] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," tech. rep., PTT Research, Leidschendam, The Netherlands, 1992. Disponible en: `https://aes2.org/publications/elibrary-page/?id=7019`.

[42] International Electrotechnical Commission, "IEC 61672-1:2013: Electroacoustics – sound level meters – part 1: Specifications," international standard, International Electrotechnical Commission, Geneva, Switzerland, 2013. Second edition.

[43] S. Nesmachnow and S. Iturriaga, "Cluster-uy: Collaborative scientific high performance computing in uruguay," in *Supercomputing. ISUM 2019* (M. Torres and J. Klapp, eds.), Communications in Computer and Information Science, Springer, Cham, 2019.

[44] A. Arimón, G. Mazzeo, and R. Torrado, "Restauración de grabaciones musicológicas mediante técnicas de denoising: Sustracción espectral y aprendizaje profundo." `https://denoisify-f204a2.pages.fing.edu.uy/`, 2025. Último acceso: 30 de noviembre de 2025.

[45] V. Ruiz, "Amalia de la vega y lauro ayestarán: estudio de las músicas en uruguay," tech. rep., Universidad de la República, Facultad de Humanidades y Ciencias de la Educación, 2019. Publicado el 8 de noviembre de 2019.

[46] Uruguay Educa, ANEP, "Estilo (música criolla)." `https://uruguayeduca.anep.edu.uy/recursos-educativos/5206`, s.f.

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

[48] R. Torrado, "Ss denoisify: A structured implementation of spectral subtraction for audio restoration." `https://github.com/RoTorrado/denoisify-spectral-subtraction`, 2025. Accedido: 30-nov-2025.

[49] ITU-T, "G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)." ITU-T Recommendation G.729, June 2012. Series G: Transmission Systems and Media, Digital Systems and Networks.

[50] ETSI, "GSM 06.42: Digital cellular telecommunications system (Phase 2+); Half rate speech; Voice Activity Detector (VAD) for half rate speech traffic channels," Tech. Rep. EN 300 973 V6.0.1, European Telecommunications Standards Institute, June 1999.

[51] I. N. de Ciberseguridad (INCIBE), "¿pueden escuchar los asistentes virtuales nuestras conversaciones?," 2023.

[52] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201–212, 1976.

[53] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 139, no. 4, pp. 377–380, 1992.

[54] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *Proceedings of the IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering (TENCON)*, (Beijing, China), pp. 321–324, October 1993.

[55] S. A. McClellan and J. D. Gibson, "Spectral entropy: An alternative indicator for rate allocation?," in *Proceedings of the IEEE International Conference on*

Acoustics, Speech, and Signal Processing (ICASSP), (Adelaide, Australia), pp. 201–204, April 1994.

[56] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, (Aalborg, Denmark), pp. 1887–1890, September 2001.

[57] K. Sakhnov, E. Verteletskaya, and B. Simak, "Dynamical energy-based speech/silence detector for speech enhancement applications," in *Proceedings of The World Congress on Engineering 2009*, pp. 801–806, 2009.

[58] E. Zenteno and M. Sotomayor, "Robust voice activity detection algorithm using spectrum estimation and dynamic thresholding," in *2009 IEEE Latin-American Conference on Communications*, pp. 1–5, 2009.

[59] A. Adjila, M. Ahfir, and D. Ziadi, "Silence detection and removal method based on the continuous average energy of speech signal," in *2021 International Conference on Information Systems and Advanced Technologies (ICISAT)*, pp. 1–5, 2021.

[60] S. Basu, "A linked-hmm model for robust voicing and speech detection," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 1, pp. I–I, 2003.

[61] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 ibm spine evaluation system," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–53–I–56, 2002.

[62] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Interspeech 2005*, pp. 369–372, 2005.

[63] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[64] Library of Congress, "Sticky-shed syndrome research." `https://www.loc.gov/preservation/scientists/projects/sticky_shed.html`, 2021. Accedido el 5 de agosto de 2025.

[65] J. W. C. Van Bogart, "General comments on the stability of videotape." `https://www.loc.gov/static/programs/national-film-preservation-board/documents/tvvanbogart.pdf`, 1996. Accedido el 5 de agosto de 2025.

[66] A. Pace, "Magnetic tape alert project report," tech. rep., International Association of Sound and Audiovisual Archives, 2020.

Índice de tablas

Índice de figuras

Esta es la última página.
Compilado el lunes 5 enero, 2026.
http://iie.fing.edu.uy/