MEJORA DE LA COMPRENSIÓN DEL HABLA EN AMBIENTES RUIDOSOS A TRAVÉS DE UN ENFOQUE DE APRENDIZAJE PROFUNDO

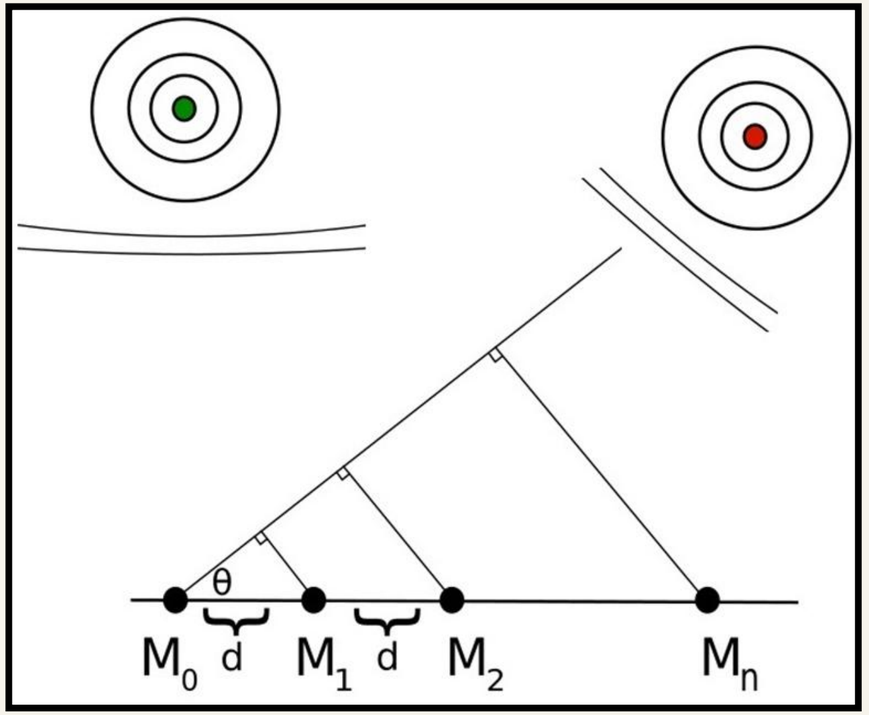
Daniel Rubén Ochoa Galván

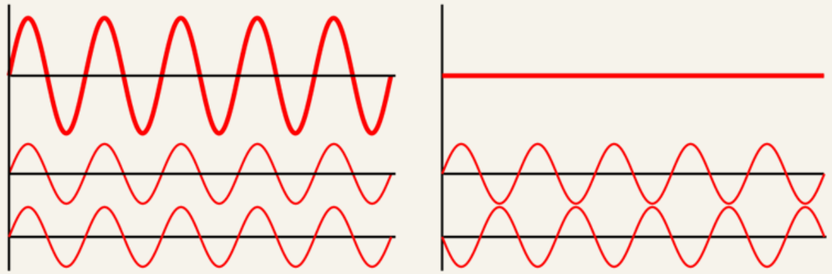
Contexto

- La **comprensión del habla** se ve mermada en entornos ruidosos, donde un **emisor** o **hablante** compite con **fuentes de ruido**.
- La **relación señal-ruido (SNR)** es una métrica que **compara** el nivel de la **señal de interés** contra el nivel del **ruido de fondo,** expresado en decibeles.

Técnicas de filtrado

- ***Beamforming:*** Técnica que utiliza más de un microfono (más de un solo canal) para captar sonido y sus fuentes en relación al **espacio** y **tiempo.**
- **Cancelación de ruido:** Manipulación del sonido a través de **interferencia** de ondas **constructivas** y **destructivas.**
- Problemas para implementar estas técnicas a la escala de **aparatos auditivos** comerciales.





Deep learning

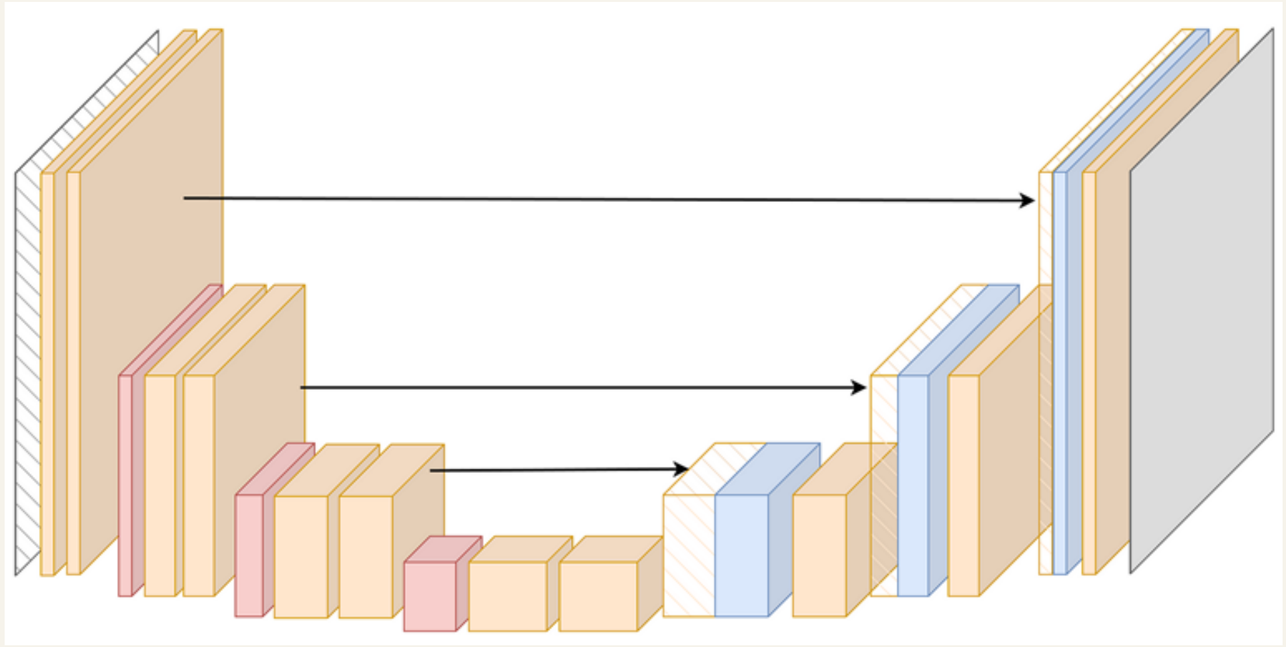
- Enfoque en utilizar el **software** como herramienta para la reducción de ruido, a diferencia de la dependencia en configuraciones particulares de **hardware.**
- Utilización de una **red neuronal** que recibe entradas de **espectrogramas** de audio en canal **mono** sobre una **persona hablando** con **ruido** de fondo, y regresa el espectrograma **limpio** con el ruido **reducido**.

Dataset de entrenamiento: Valentini

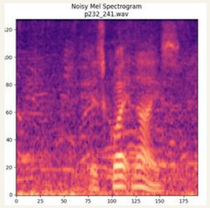
- Un conjunto de datos de archivos *.wav* para **entrenamiento y pruebas**, con ficheros de audio en canal **mono** de hombres y mujeres hablando, separado en condiciones **limpias** y condiciones de **ruido**, operando a **48kHz**.
- El conjunto de datos ruidoso fue creado con 10 tipos de ruido (2 **artificiales** y 8 **reales** obtenidas de la base de datos ***Demand***) con valores variables de SNR de 15, 10, 5 y 0 decibeles.

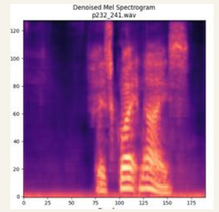
Arquitectura: Red neuronal U-Net

- Una red neuronal **convolucional** ampliamente utilizado en segmentación de imágenes, utilizando un **codificador** (*encoder*) para *downsampling* con capas convolucionales, y un **descifrador** (*decoder*) para *upsampling* con convoluciones transpuestas.



Arquitectura: Red neuronal U-Net





Entrada: Espectrograma de Mel con ruido

Conv(16) x2 + MaxPool

Conv(32) x2 + MaxPool

Conv(64) x2 + MaxPool

Conv(128) x2

Bottleneck

UpConv(64) + merge + Conv(64)×2

UpConv(32) + merge + Conv(32)×2

UpConv(16) + merge + Conv(16)×2

Salida: Espectrograma de Mel limpio

Preprocesamiento

- Carga de los ficheros limpios y ruidosos *.wav*
  - Se submuestrea de 48 kHz a 16 kHz

- Para cada par de audio ruidoso-limpio correspondientes, se calcula la **transformada de Fourier de tiempo reducido (STFTs).**
  - Desmenuza la señal en **segmentos de tiempo cortos traslapados,** aplicando una función de ventana **Hanning** para aislar cada segmento.
  - Se realiza una transformada de Fourier en el segmento para obtener sus **frecuencias.**
  - Mueve la ventana hacia adelante, utilizando un *hop_length* definido, para producir una representación de **tiempo-frecuencia** de la señal.

Preprocesamiento

Habla con ruido

$$y[n] = x[n] + z[n]$$

Método OLA (overlap-and-add): División de fotogramos de cierta longitud, traslapadas con sus fotogramas adyacentes

$$y_w^{(m)}[n] = y[n + mL_s]w[n] \text{ for } \begin{matrix} 0 \le n \le L_f - 1; \\ 0 \le m \le M - 1. \end{matrix}$$

$$= y[i]w[i - mL_s] \text{ for } \begin{matrix} i = n + mL_s; \\ 0 \le i \le (M-1)L_s + L_f - 1. \end{matrix}$$

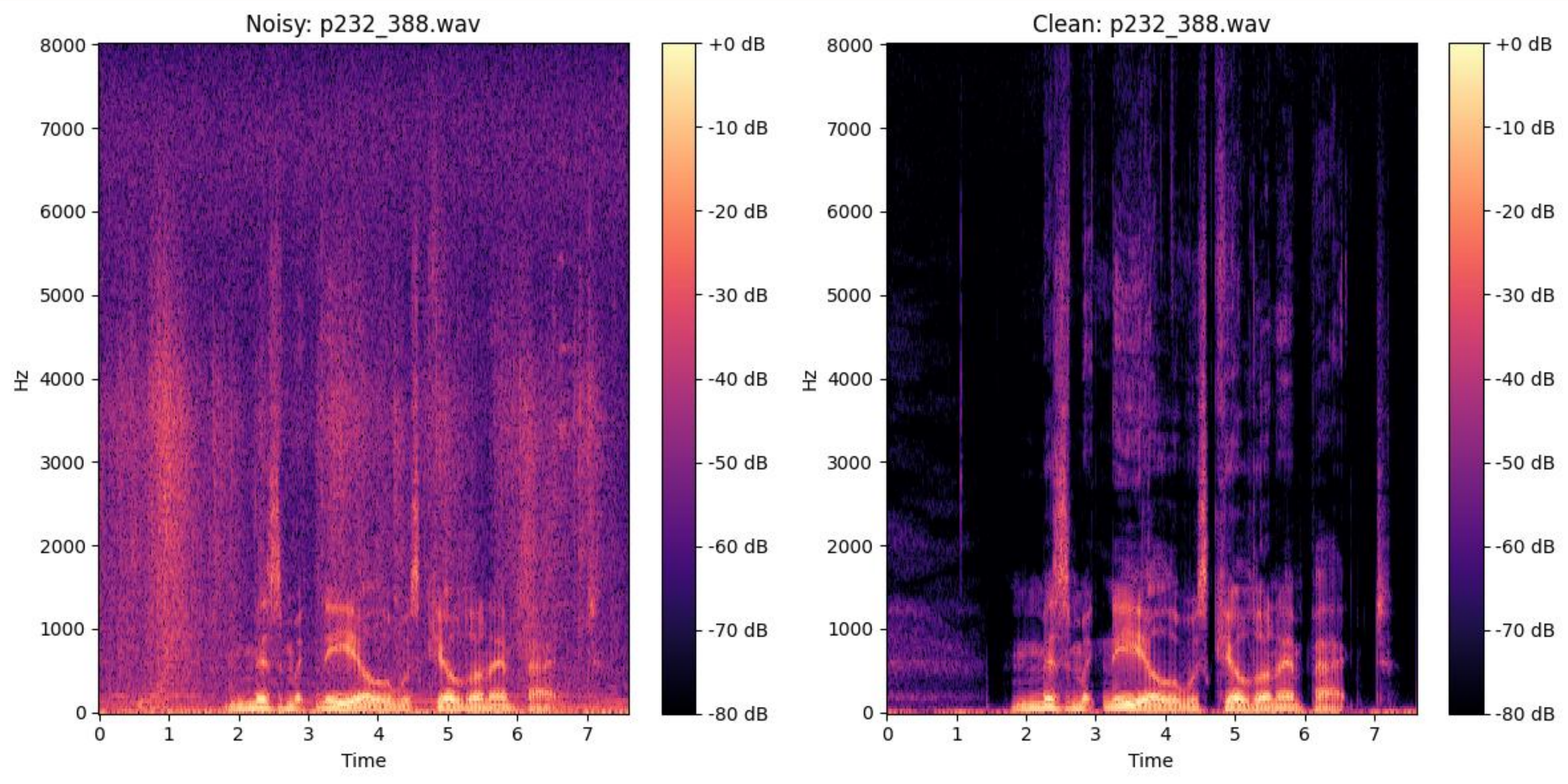
Ventana de Hamming

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L_f}\right), & n = 0, 1, \cdots, L_f - 1; \\ 0, & \text{otherwise.} \end{cases}$$

Transformada de Fourier de tiempo reducido (STFT)

$$\Upsilon_{DFT}^{(m)}[k] = \sum_{n=0}^{L_f-1} y_w^{(m)}[n] e^{-i\frac{2\pi}{L_f}kn}, \; k = 0, 1, 2, \cdots, L_f - 1.$$

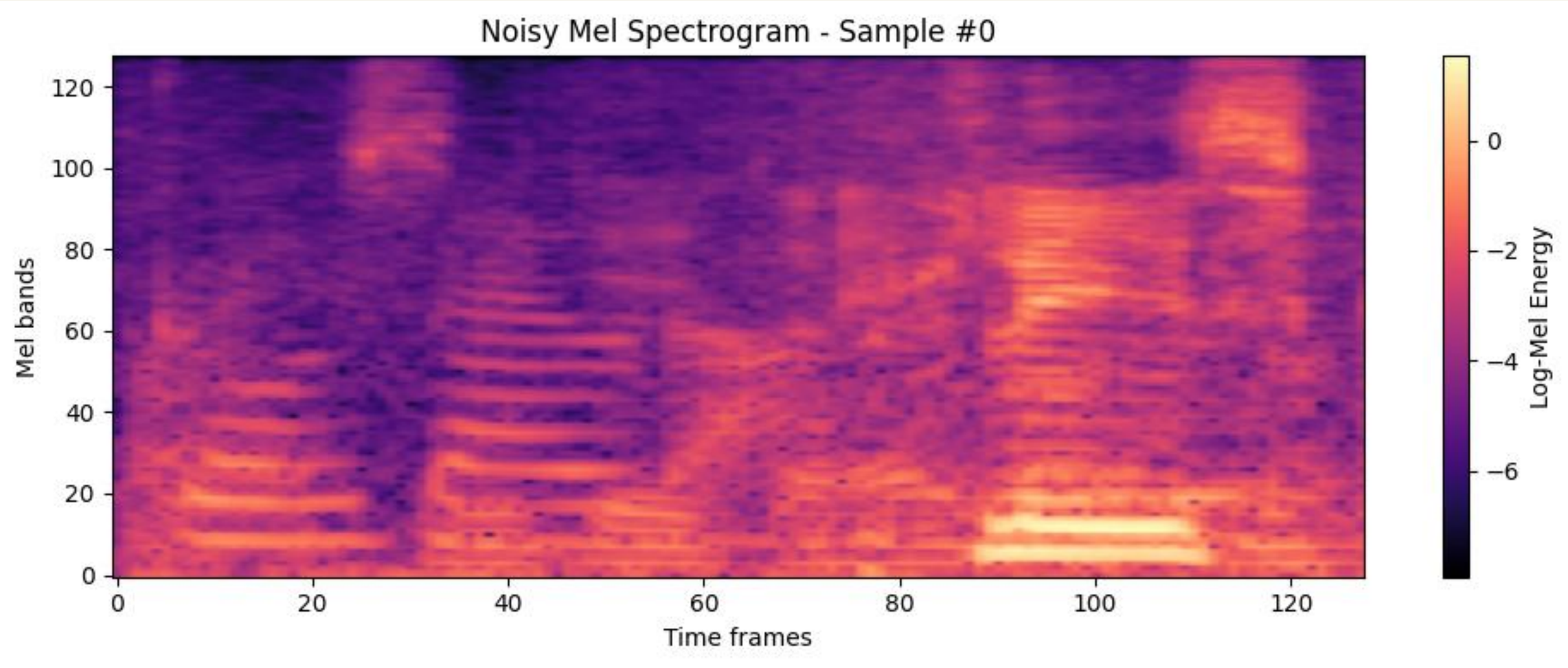
STFTs de una señal



Preprocesamiento

- Se calcula el espectrograma de **potencia**, convirtiendo los valores complejos del STFT en información de **magnitud,** enfatizándola y despreciando la **fase.**
- Se aplica un filtro de **Mel,** escalando el espectrograma comprimiendo las frecuencias altas y esparciendo las bajas, para asemejar cómo los humanos escuchan la frecuencia.
- Se escala el espectrograma de Mel de forma **logarítmica** (como los humanos perciben el volumen), para comprimir el rango dinámico.

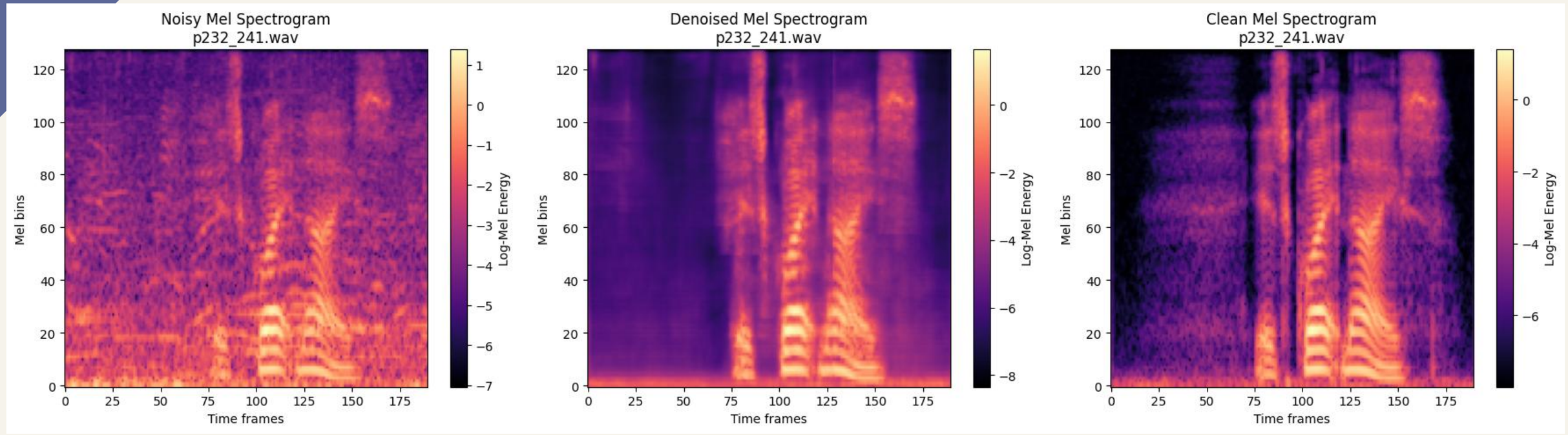
Espectrograma de Mel



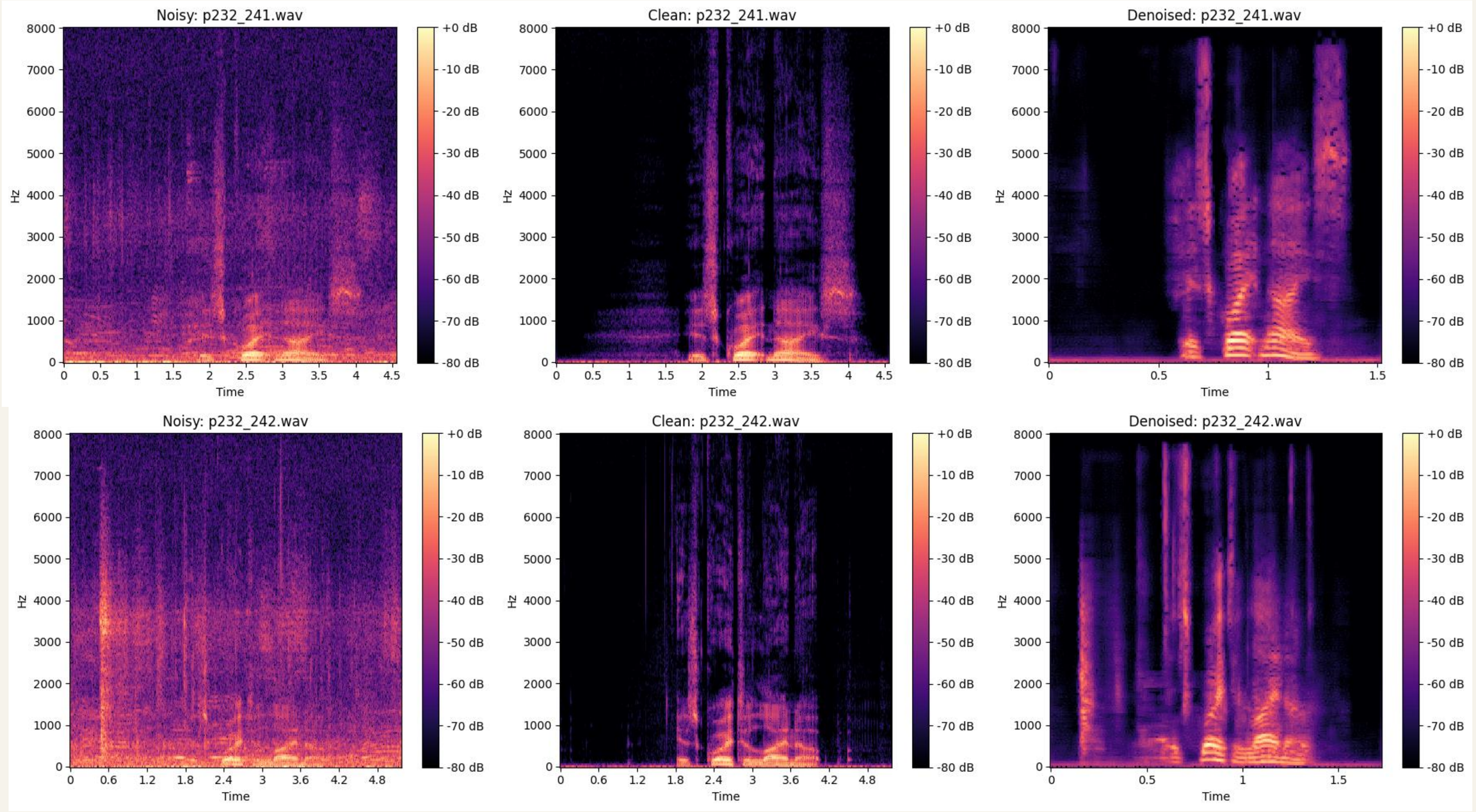
Posprocesamiento

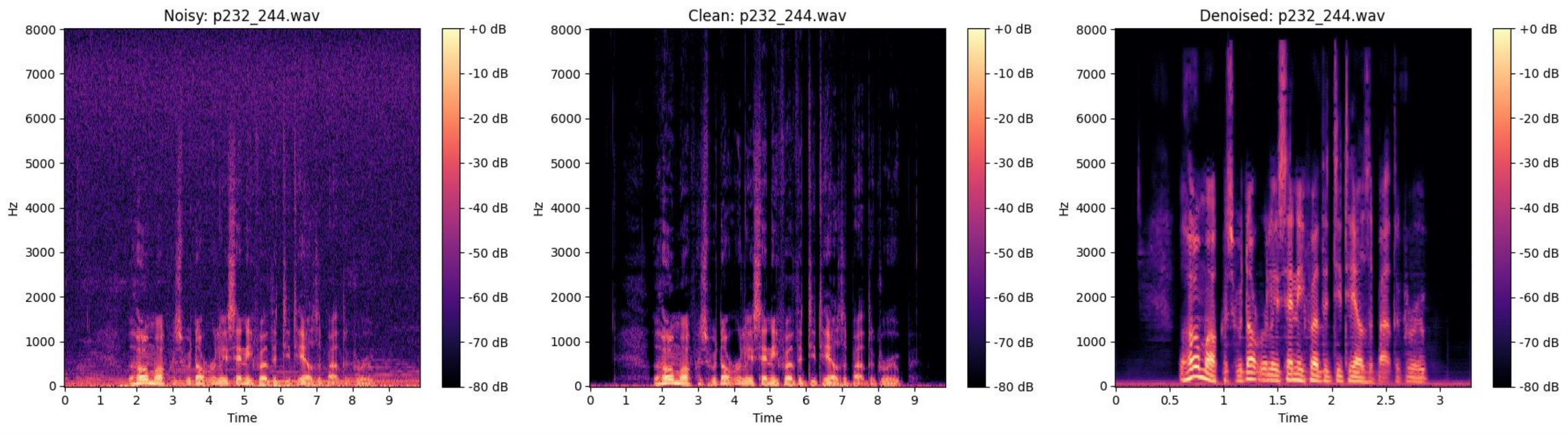
- Con el espectrograma arrojado por la red neuronal, se transforma a la inversa para regresar al formato original
- Se invierte la escala logarítmica
- Se invierte el escalado de Mel, obteniendo el STFT.
- Se utiliza el algoritmo de **Griffin-Lim** (GLA), un método de reconstrucción que recupera una **estimación de la fase** (a este punto muy ruidoso o no disponible), en base a la **información que comparten los fotogramas traslapados**, contenidas en el espectro de magnitud. Esta información tiene componentes de **magnitud** y **fase**, y se usa para estimar la fase, **reconstruyendo la señal de audio**.

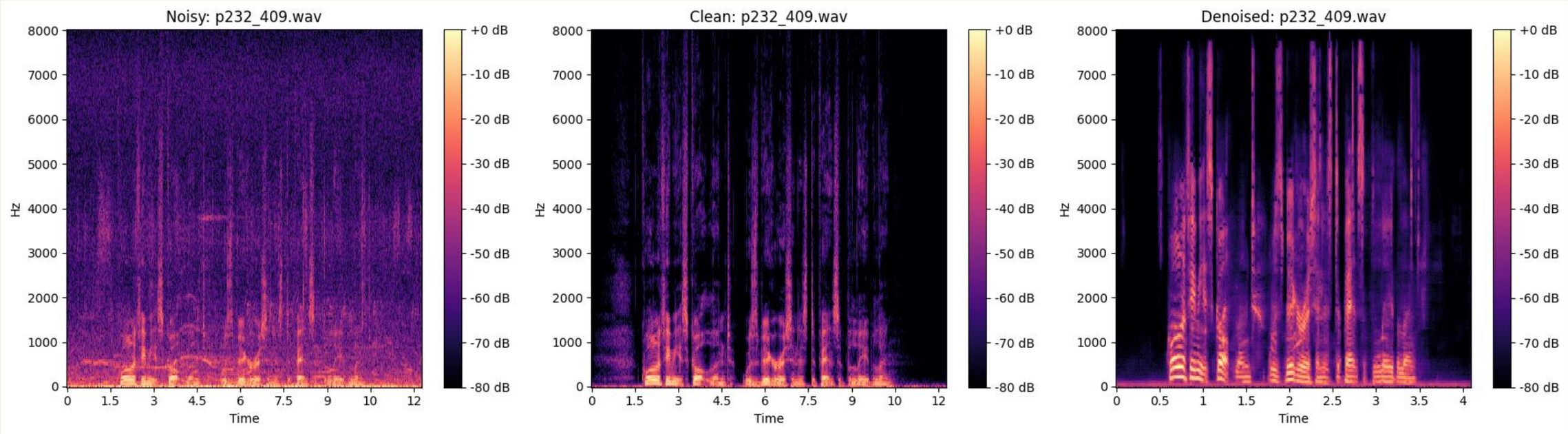
Resultados

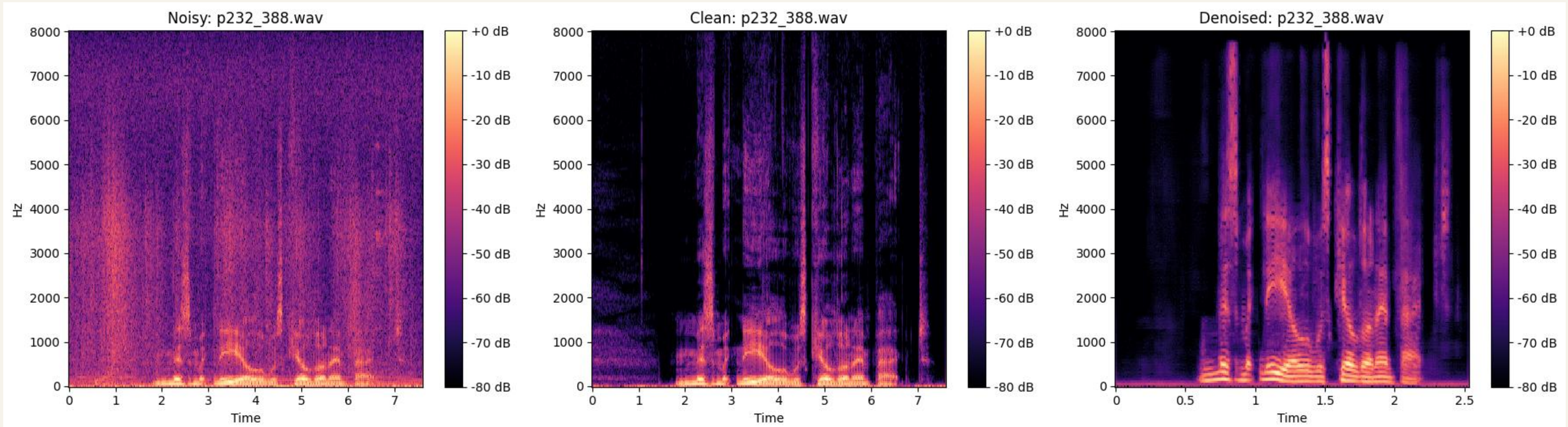

Noisy Mel Spectrogram
p232_241.wav
Denoised Mel Spectrogram
p232_241.wav
Clean Mel Spectrogram
p232_241.wav
Mel bins
Time frames
Log-Mel Energy

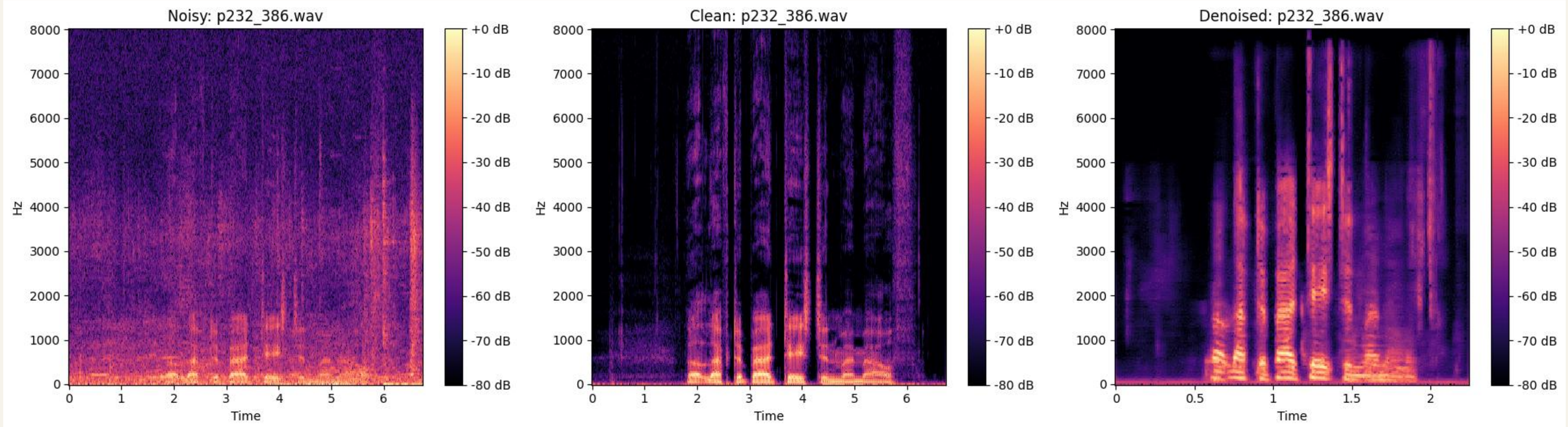
STFTs

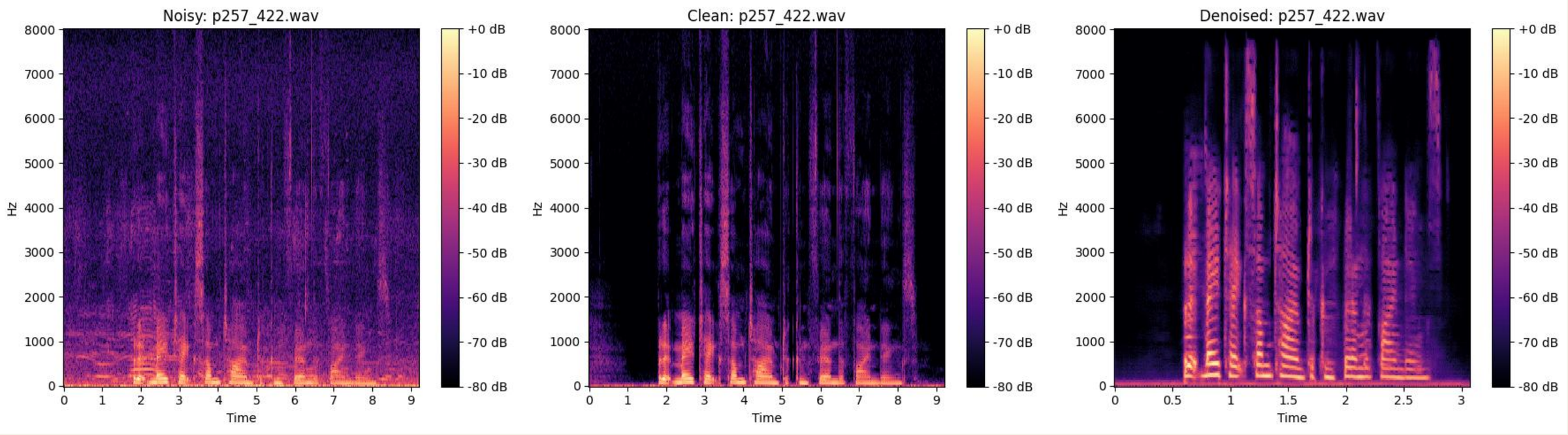

Noisy: p232_241.wav
Clean: p232_241.wav
Denoised: p232_241.wav
Noisy: p232_242.wav
Clean: p232_242.wav
Denoised: p232_242.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

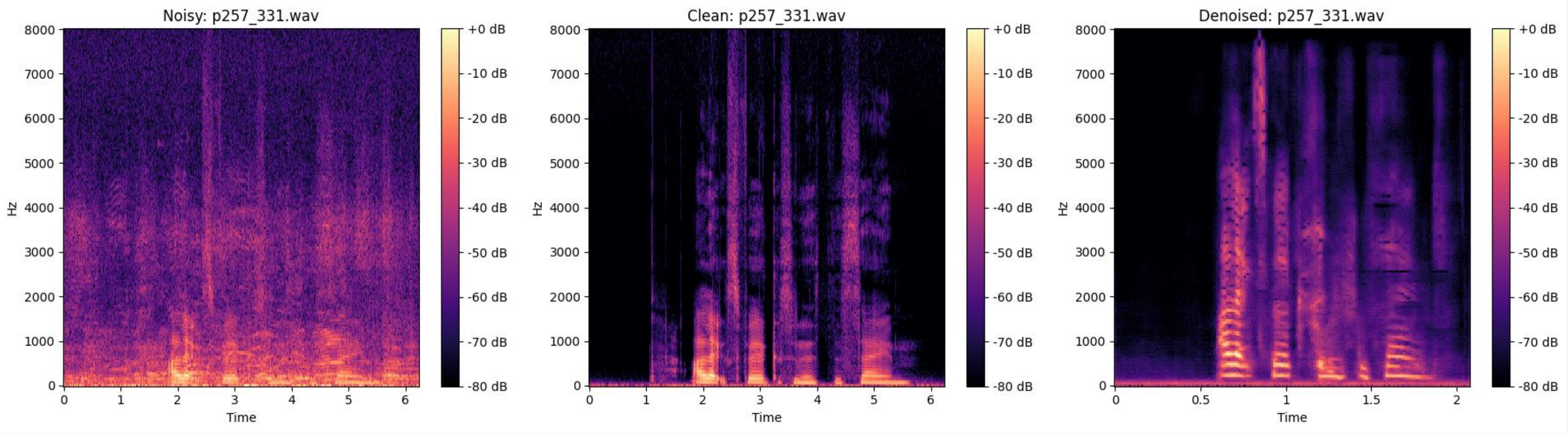

Noisy: p232_244.wav
Clean: p232_244.wav
Denoised: p232_244.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

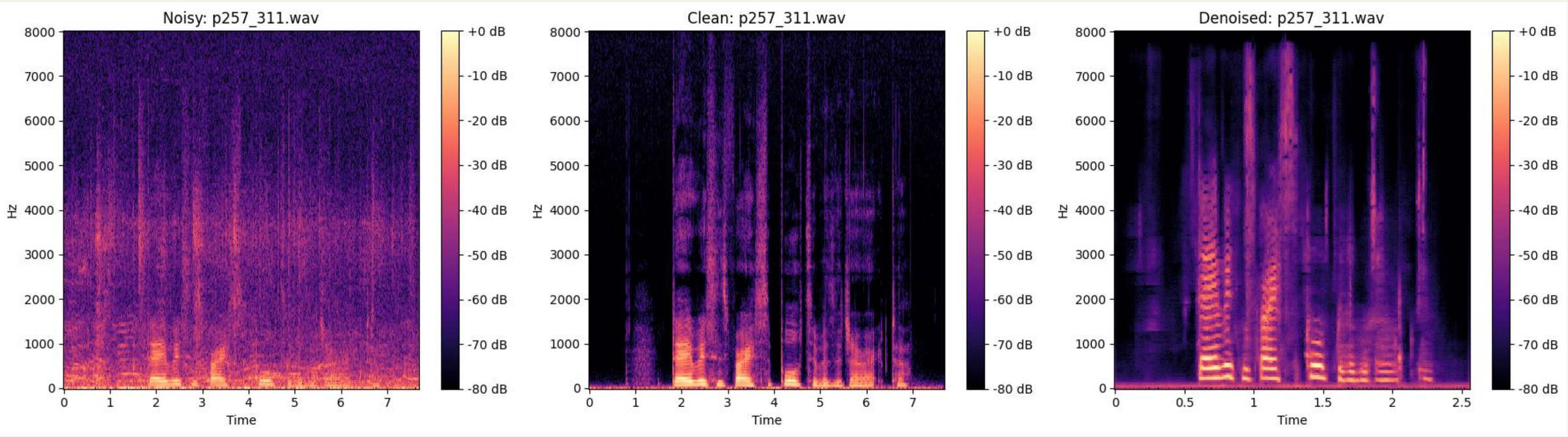

Noisy: p232_409.wav
Clean: p232_409.wav
Denoised: p232_409.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

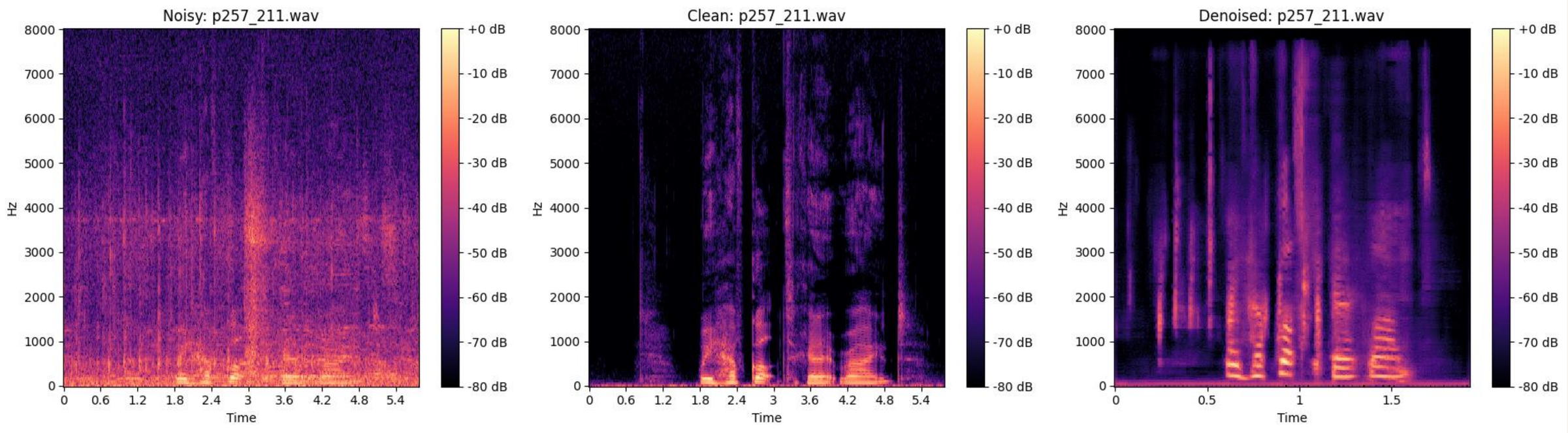

Noisy: p232_388.wav
Clean: p232_388.wav
Denoised: p232_388.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

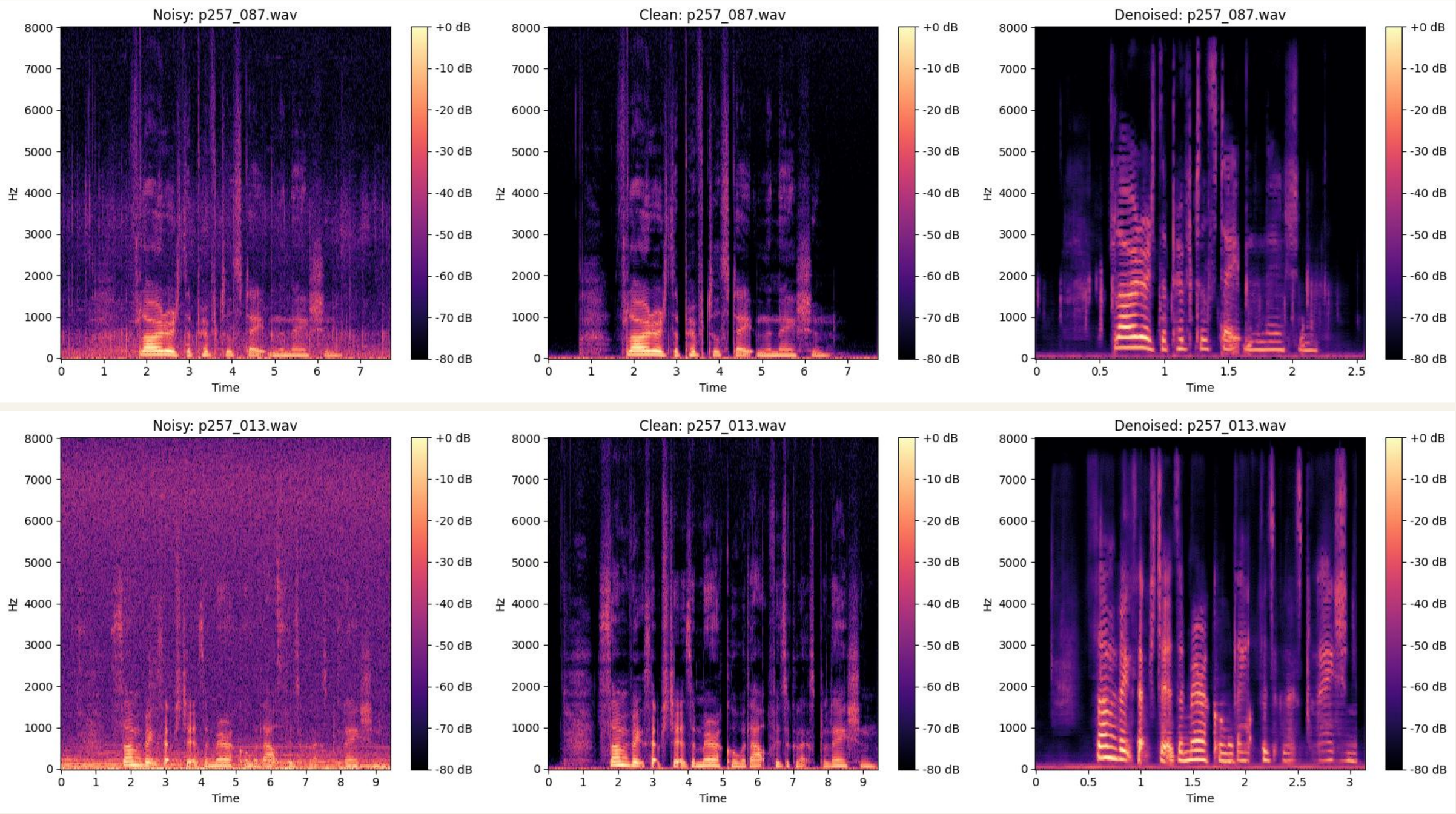

Noisy: p232_386.wav
Clean: p232_386.wav
Denoised: p232_386.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Noisy: p257_422.wav
Clean: p257_422.wav
Denoised: p257_422.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Noisy: p257_331.wav
Clean: p257_331.wav
Denoised: p257_331.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Noisy: p257_311.wav
Clean: p257_311.wav
Denoised: p257_311.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Noisy: p257_211.wav
Clean: p257_211.wav
Denoised: p257_211.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Noisy: p257_087.wav
Clean: p257_087.wav
Denoised: p257_087.wav
Noisy: p257_013.wav
Clean: p257_013.wav
Denoised: p257_013.wav
Hz
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

Referencias

- Bulut, A. E., & Koishida, K. (2020, May). Low-latency single channel speech enhancement using u-net convolutional neural networks. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6214-6218). IEEE.
- Diehl, P. U., Singer, Y., Zilly, H., Schönfeld, U., Meyer-Rachner, P., Berry, M., ... & Hofmann, V. M. (2023). Restoring speech intelligibility for hearing aid users with deep learning. Scientific Reports, 13(1), 2719.
- Gul, S., & Khan, M. S. (2023). A survey of audio enhancement algorithms for music, speech, bioacoustics, biomedical, industrial, and environmental sounds by image U-Net. IEEE Access, 11, 144456-144483.
- Hu, H. T., & Lee, T. T. (2024). Options for Performing DNN-Based Causal Speech Denoising Using the U-Net Architecture. Applied System Innovation, 7(6), 120.
- Mukherjee, A., Banerjee, R., & Ghose, A. (2023). A novel U-Net architecture for denoising of real-world noise corrupted phonocardiogram signal. arXiv preprint arXiv:2310.00216.
- Nustede, E. J., & Anemüller, J. (2021, August). Towards speech enhancement using a variational U-Net architecture. In 2021 29th European Signal Processing Conference (EUSIPCO) (pp. 481-485). IEEE.