PRÁCTICA 4

Compresión de Audio – MPEG 1

Universidad Pontifica de
Comillas ICAI

Andrés Sánchez de Ágreda – 202108563
Servicios Telemáticos Multimedia

ÍNDICE

1. Introducción

Esta práctica analiza la compresión de audio desde dos frentes complementarios. En la Parte 1 se estudia un fragmento de piano reverberante con el modelo psicoacústico de MPEG-1 Layer I: detección de tonales/no tonales, umbrales de enmascaramiento (Bark y subbandas) y SMR. En la Parte 2 se convierte el WAV→MP3 a distintos bitrates (y con Audacity), comparando tamaños y evaluando la calidad mediante métricas objetivas (SNR, SegSNR, RMSE, LSD) y escucha, para relacionar teoría psicoacústica y rendimiento real del códec.Análisis MPEG1 audio en nivel 1

2. Fase de Análisis MPEG1 audio en nivel 1

2.1. Elegir un fichero de audio (no voz) en formato .wav. Indicar qué tipo de audio se ha elegido y el tamaño del mismo.

Se ha elegido el archivo de audio *P4audioSTM.wav*, una canción de producción propia de música clásica. Este audio tiene dos canales (estéreo) y está muestreado a 44,1 kHz. Tiene una duración de 19,97 segundos y un peso de 5,04 MB. El tamaño de este archivo coincide con el típico de PCM 24-bit.

2.2. Usar el código de MPEG1 nivel 1 para tratar el fichero de audio. Encontrar las componentes tonales y no tonales en un pequeño fragmento del fichero de audio.

Se ha seleccionado el fragmento: t∈[5.00,5.05] s (N_frame=2048, N_FFT=8192). Este audio se trata de una pista de piano con bastante reverberación. Predomina un fondo difuso con pocos picos, lo que sugiere umbrales de enmascaramiento altos y SMR más ajustada en graves/medios.

Se han detectado los siguientes componentes tonales y no tonales en el audio:

#	Frecuencia (Hz)	Nivel L (dB rel.)
1	7100.6	-100.0
2	1189.7	-44.0
3	839.8	-26.6
4	6562.2	-99.0
5	21414.8	-112.4
6	2761.6	-79.0
7	19256.1	-114.6
8	4915.0	-89.3

Tabla 1: Componentes tonales detectadas

#	f_rep (Hz)	Nivel de banda (dB rel.)	CB
1	85.3	-0.8	1
2	108.4	-7.8	2
3	266.9	-22.6	3

4	341.1	-16.1	4
5	476.1	-19.5	5
6	576.0	-17.6	6
7	725.1	-5.8	7
8	916.3	-41.0	8
9	991.0	-25.0	9
10	1091.9	-40.3	10
11	1331.6	-30.7	11
12	1578.1	-33.9	12
13	1914.5	-42.0	13
14	2373.7	-37.6	15
15	3203.5	-42.4	17

Tabla 2: Componentes no tonales detectados

A continuación, se han incluido dos gráficas que permiten visualizar con más detalle estás componentes y su situación en las bandas.

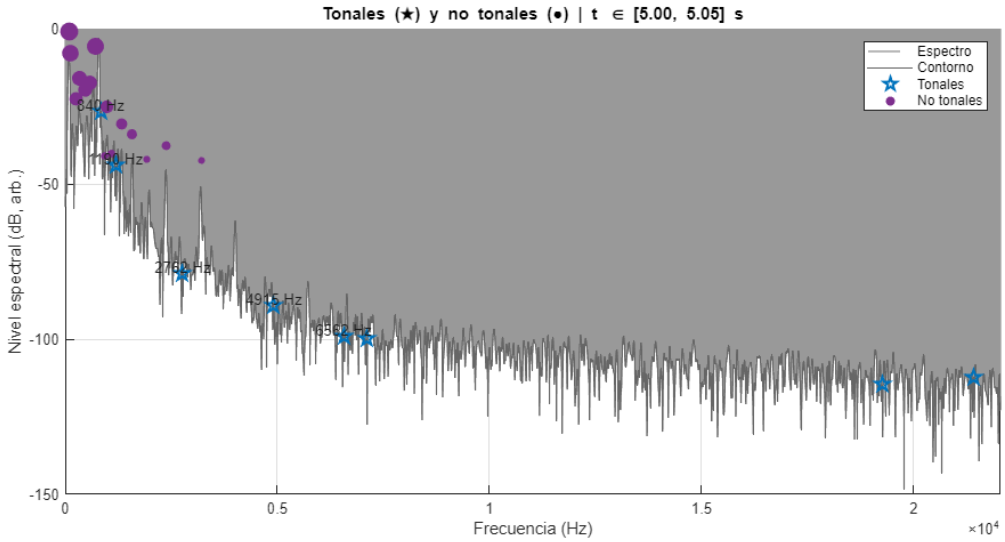

Ilustración 1: Espectro con marcadores

En la figura, las estrellas señalan las componentes tonales y los círculos morados las no tonales (su tamaño refleja el nivel). El espectro muestra una pendiente descendente y un fondo difuso de energía no tonal en graves/medios debido a la reverberación. Destacan picos en torno a 840 Hz y 1.19 kHz, asociados a parciales armónicos del acorde o la pulsación del piano.

En la figura, las estrellas señalan las componentes tonales y los círculos morados las no tonales (su tamaño refleja el nivel). El espectro muestra una pendiente descendente y un fondo difuso de energía no tonal en graves/medios debido a la reverberación. Destacan picos en torno a 840 Hz y 1.19 kHz, asociados a parciales armónicos del acorde o la pulsación del piano.

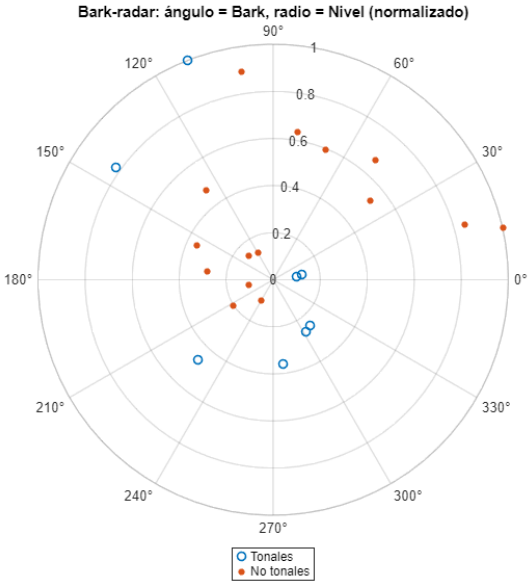

Ilustración 2: Radar Bark

En el diagrama polar, el ángulo indica la posición en Bark y el radio el nivel normalizado. Los no tonales se concentran en Bark bajos–medios, mientras que los tonales aparecen más dispersos (pocos, pero dominantes donde surgen), coherente con un timbre sostenido por la reverberación.

El fragmento (piano con fuerte reverberación) muestra en las gráficas 8 tonales poco numerosos, con picos claros en torno a 840 Hz y 1.19 kHz, y una base no tonal amplia concentrada en 80–750 Hz que "alfombra" los graves/medios y genera una pendiente descendente del espectro.

En conjunto, el sonido resulta de pocos picos dominantes sobre un fondo difuso reverberante. Como conclusión del apartado, estos resultados anticipan para (c)–(e) umbrales de enmascaramiento elevados en las primeras bandas críticas y, por tanto, SMR más ajustadas en graves/medios que, en los agudos, donde la energía y los tonales son menores.

2.3. Representar las componentes de enmascaramiento y sus umbrales para dos pequeños fragmentos del fichero de audio

Para este apartado se han calculado las curvas de enmascaramiento y el umbral global T_g con niveles normalizados al pico de cada fragmento. Fragmento A: $t \in [5.00,5.05]$s, 8 tonales y 23 no tonales. Fragmento B: $t \in [12.50,12.55]$s, 1 tonal y 24 no tonales.

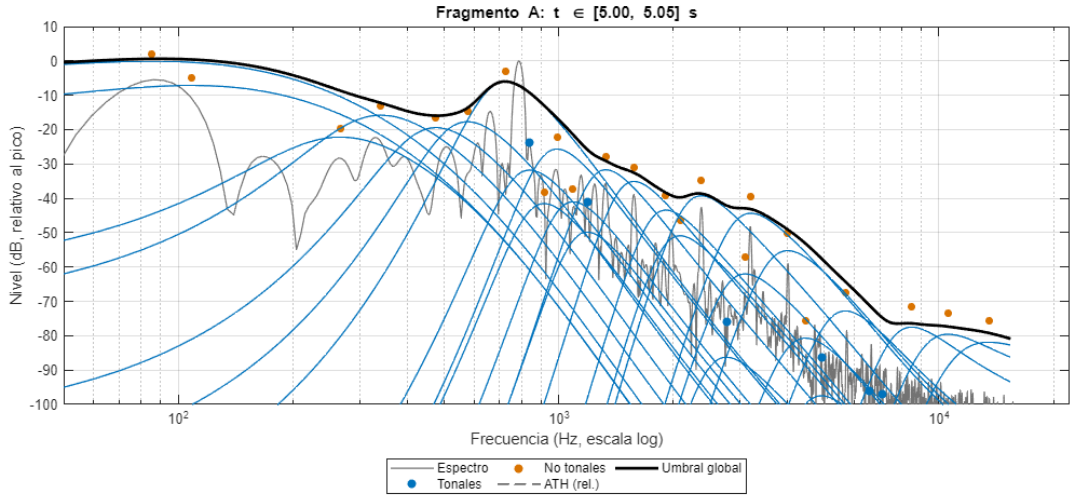

Ilustración 4: Componentes y umbral del fragmento A

En el espectro aparece una pendiente descendente sobre la que se superponen las curvas individuales de enmascaramiento; en torno a 0.8–1.2 kHz los picos tonales elevan localmente el umbral. El T_g(negro) sigue la envolvente de los mascaradores y permanece por debajo del espectro; el ATH relativo queda bastante más bajo y sólo domina donde faltan enmascaradores. El patrón refleja un piano con parciales definidos que, sumados a la reverberación, generan enmascaramiento marcado en bandas medias.

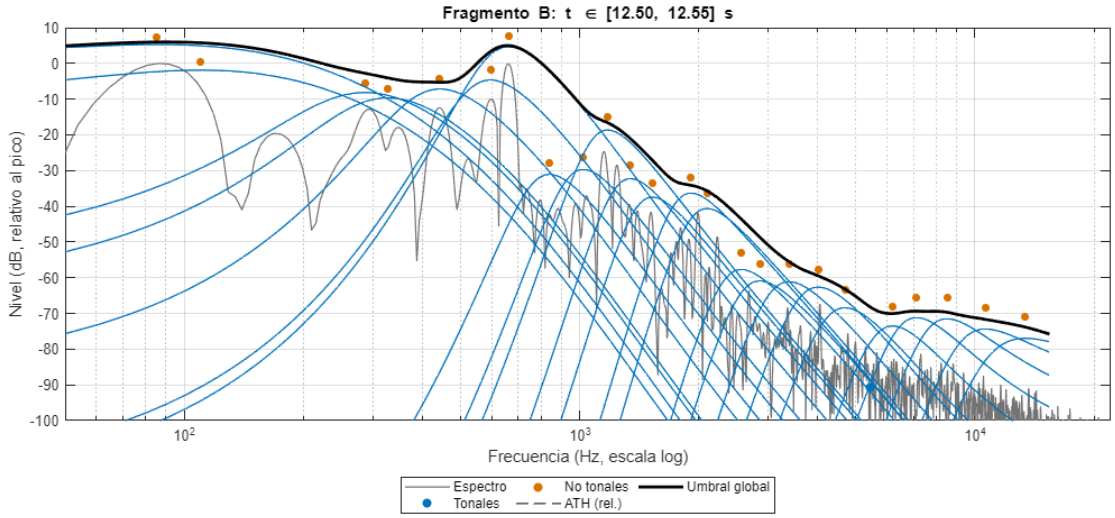

Ilustración 3: Componentes y umbral del fragmento B

En este espectro la energía está más difusa y concentrada en graves-medios, con menos picos definidos. El T_g resulta más continuo y se mantiene relativamente alto hasta ~1 kHz, descendiendo después hacia las altas frecuencias; de nuevo queda por debajo del espectro y por encima del ATH relativo. El resultado es coherente con una pulsación menos rica en parciales dominantes y una reverberación que “alfombra” las primeras bandas críticas.

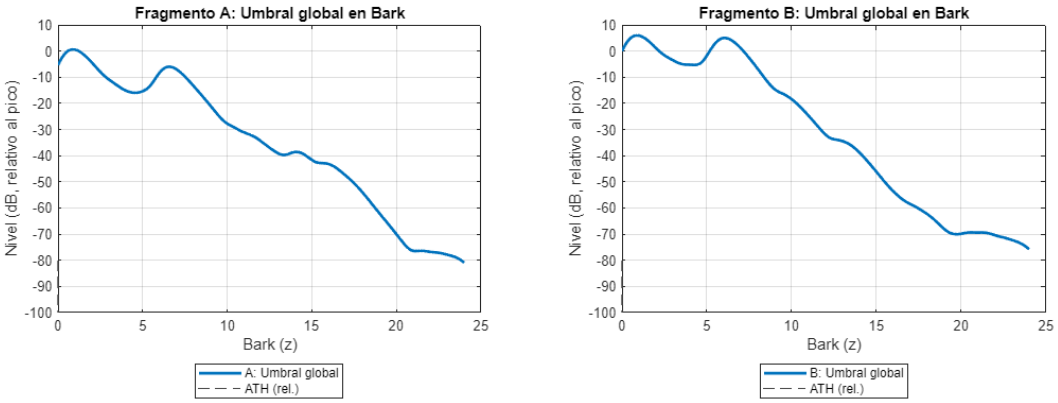

Ilustración 5: Comparativa en escala Bark

Ambos $T_g(z)$decrecen con el índice Bark. En **A** el umbral cae algo más en bandas altas (≈ −85 dB al final), mientras que en **B** permanece unos decibelios por encima (≈ −75 dB), indicio de mayor densidad difusa en ese fragmento.

El fragmento A está más influido por tonales localizados (parciales del piano) y el B por un enmascaramiento homogéneo de graves-medios. Estas diferencias anticipan umbrales más altos y SMR más ajustada en bajas-medias frecuencias, especialmente en B, y mejor separación en algunas sub-bandas de medios-altos en A.

2.4. Representar el umbral de enmascaramiento mínimo en cada sub-banda para dos pequeños fragmentos del fichero de audio

En el cuarto apartado se proyectó el umbral global $T_g(f)$de cada fragmento sobre las 32 sub-bandas uniformes de Layer I en $[0, 22050]$Hz, tomando en cada una el mínimo $T_{\min}$. Los niveles están normalizados al pico de cada fragmento.

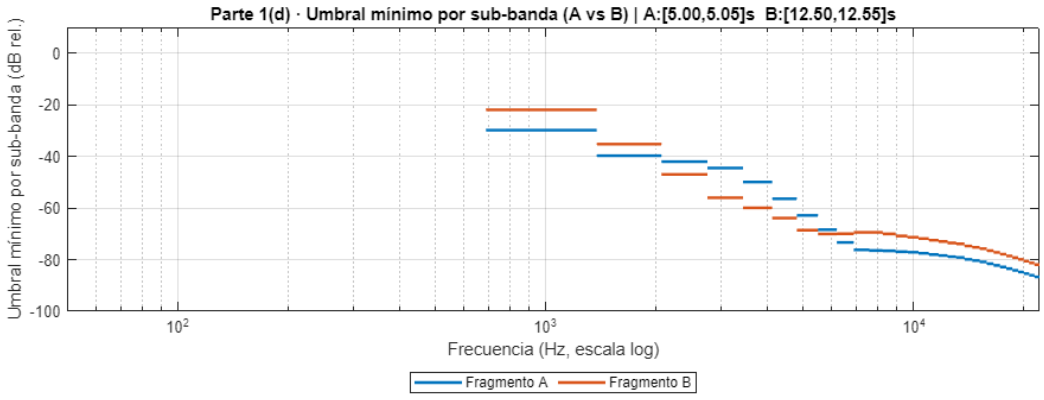

Ilustración 6: Umbral mínimo por sub-banda

En el fragmento A [5.00,5.05]s el umbral mínimo presenta una caída progresiva con la frecuencia: desde alrededor de −20 dB en las primeras subbandas hasta valores próximos a −90 dB en las altas. La curva en escalones y las barras muestran umbrales algo más bajos que en B a partir de ~2 kHz, indicio de mayor "respiro" en agudos. En el fragmento B [12.50,12.55]s los escalones iniciales se sitúan más altos (≈ −6…−25 dB en las primeras bandas), coherente con la mayor densidad difusa en graves/medios observada en (c), y descienden después de forma sostenida.

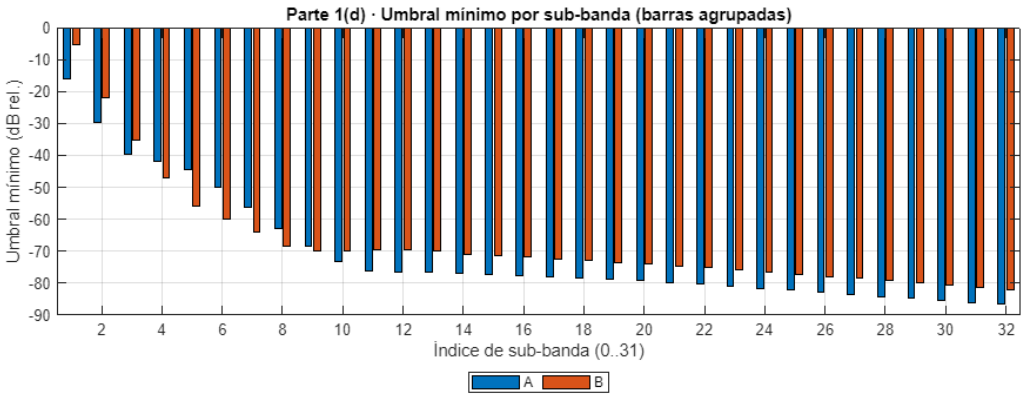

Ilustración 8: Comparación por índice de sub-banda

La representación en barras agrupadas confirma el patrón: en sub-bandas bajas (0–7) el fragmento B mantiene $T_{\min}$más alto (enmascaramiento mayor), mientras que a partir de medios-altos el fragmento A tiende a quedar ligeramente por debajo, lo que favorece una codificación más laxa en esas bandas para A.

El heatmap sintetiza la tendencia: región cálida (umbrales altos) al inicio de B, que se enfría con la frecuencia; en A, la transición a valores bajos llega antes y se mantiene estable en medios-altos.

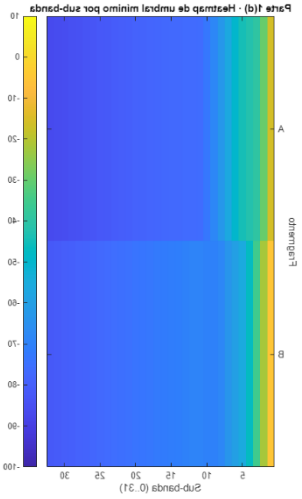

Ilustración 7: HeatMap

Por resumir un poco, la información de las sub-bandas es la siguiente:
- Fragmento A [5.00,5.05]s: media $T_{\min}$= −70.5 dB, mín = −86.7 dB, máx = −15.9 dB.

- Fragmento B [12.50,12.55]s: media $T_{\min}$= −67.3 dB, mín = −81.9 dB, máx = −5.3 dB.

En conclusión, El fragmento B presenta enmascaramiento más fuerte en las primeras sub-bandas (umbrales mínimos más altos), coherente con la reverberación dominante en graves/medios; el fragmento A muestra umbrales más bajos a partir de medios-altos, favoreciendo mayor margen de cuantización en esas frecuencias. Estas diferencias condicionan las SMR del apartado (e), que serán previsiblemente más ajustadas en B en bajas-medias frecuencias y algo más holgadas en A en la zona de agudos.

2.5. Calcular la ratio señal/máscara para dos pequeños fragmentos del fichero de audio

En este apartado se ha estimado, para las 32 sub-bandas de MPEG-1 Layer I, la SMR como diferencia entre el nivel de señal de banda y el umbral mínimo de enmascaramiento obtenido en (d). Todos los niveles están normalizados al pico de cada fragmento. Los resultados obtenidos han sido:

- En el fragmento A la SMR media es 0,8 dB, con un mínimo de −6,0 dB en la sub-banda 6.
- En el fragmento B la SMR media es −0,5 dB, con un mínimo de −6,6 dB en la sub-banda 21.

Estos valores indican que, en promedio, A mantiene un margen ligeramente positivo frente al umbral, mientras que B queda levemente por debajo en conjunto.

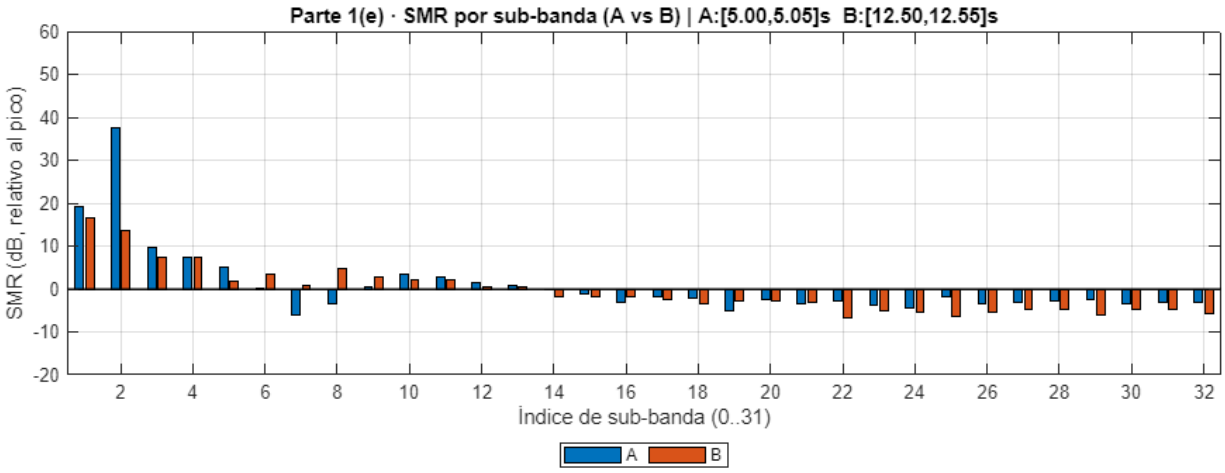

Ilustración 9: SMR por sub-banda

La representación en barras muestra que las primeras sub-bandas concentran los márgenes más altos (hasta decenas de dB en A), reflejando la energía del piano en graves inmediatos a la fundamental y los primeros parciales. A partir de medios-altos la SMR cae y muchas bandas se sitúan en torno a 0 dB o negativas, más acusado en B, consistente con su mayor enmascaramiento en esas zonas.

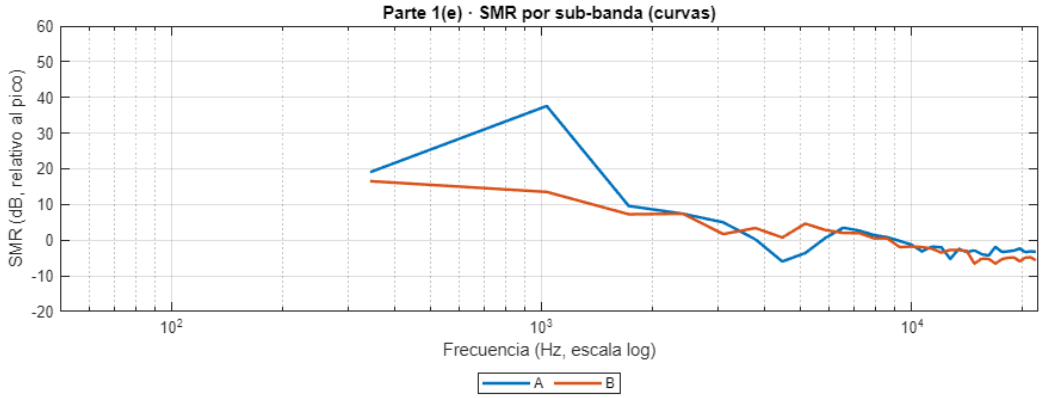

Ilustración 10: Curva de la SMR por sub-banda

La curva confirma el patrón: en A la SMR alcanza un máximo en ~1 kHz y desciende de forma casi monótona hacia altas frecuencias, cruzando 0 dB en la zona de 2–4 kHz y volviéndose negativa en agudos. En B la SMR es más contenida desde bajas frecuencias y cae antes por debajo de 0 dB, con mínimos en el rango 10–15 kHz donde el contenido útil es escaso y el umbral permanece relativamente alto por la energía difusa.

Para finalizar con la parte de nivel I, se puede concluir que el fragmento A ofrece más margen de codificación en las primeras sub-bandas y mantiene SMR ligeramente positiva en promedio; el fragmento B presenta enmascaramiento más fuerte en medias-altas y SMR global menor, lo que sugiere que, a igual tasa, B sería más sensible a artefactos en esas bandas y requeriría una cuantización más cuidadosa especialmente desde ~2 kHz hacia arriba.

3. Fase de Análisis MPEG1 audio en nivel 3

3.1. Usar el fichero de audio anterior y el código para convertir el fichero wav a mp3 usando al menos dos velocidades (tasas de bits diferentes). Indicar el tiempo transcurrido en el proceso de paso a mp3 en ambas ocasiones

Para realizar la codificación del archivo WAV a MP3 se ha utiliza el encoder FFmpeg 8.0. Se han evaluado las siguientes tasas: 320 kbps y 128 kbps (CBR). A cada codificación se midió el tiempo transcurrido y el tamaño final.

Bitrate	Tiempo (s)	Tamaño (MB)	Fichero de salida
320 kbps	1.217	0.8	...\mp3_out\P2_320kbps.mp3
128 kbps	0.381	0.3	...\mp3_out\P2_128kbps.mp3

Tabla 3: Resumen de la compresión a MP3

En resumen, La conversión a 128 kbps es ~3,2× más rápida que a 320 kbps y produce un fichero ~0,3 MB, mientras que a 320 kbps el fichero queda en ~0,8 MB con un tiempo de ~1,22 s. Estas salidas se usarán en (c)–(f) para comparar tamaños/ratios, calidad objetiva, y evaluación auditiva.

3.2. Comparar tamaños de fichero wav y los dos obtenidos al pasarlos por el compresor (ratios de compresión)

Fichero	Tamaño (MB)	Ratio (WAV/MP3)	Reducción
P2_320kbps.mp3	0.76	6.59 : 1	84.8 %
P2_128kbps.mp3	0.31	16.48 : 1	93.9 %

Tabla 4: Tamaños y ratios de compresión

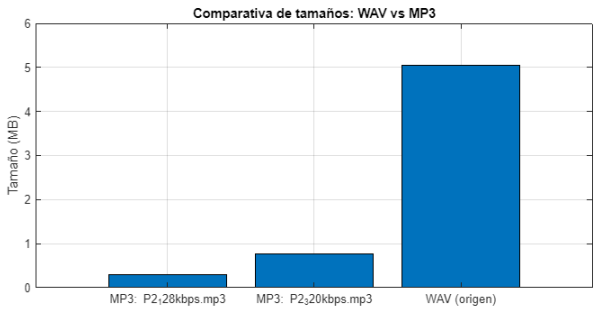

Ilustración 11: Comparativa de tamaños

El 128 kbps alcanza la mayor eficiencia (~16×), mientras que 320 kbps conserva más calidad con un tamaño ~6.6× menor que el WAV.

Teniendo en cuenta estos datos, si se prioriza eficiencia, 128 kbps es claramente superior; si se busca calidad con un peso aún muy bajo, 320 kbps ofrece un buen compromiso.

3.3. Comentar la calidad obtenida en todos los casos. Evaluar diferencia o errores de manera cuantitativa

Al reproducir los audios con el altavoz integrado los tres suenan muy similares; en 128 kbps se aprecia algo menos de profundidad (ligera pérdida de "aire" en agudos y colas de reverb).

Fichero	SNR (dB)	SegSNR (dB)	RMSE	LSD (dB)	Lag (muestras)
P2_128kbps.mp3	38.60	34.95	0.0006	2.07	1729

P2_320kbps.mp3	39.57	34.96	0.0005	1.58	1729

Ilustración 12: Tabla comparativa de archivos MpP3

320 kbps mejora ligeramente SNR/RMSE y reduce la distancia log-espectral (LSD); la SegSNR es prácticamente igual (el material es estable y la medida satura alrededor de 35 dB). El lag común corresponde al retardo de códec/padding.

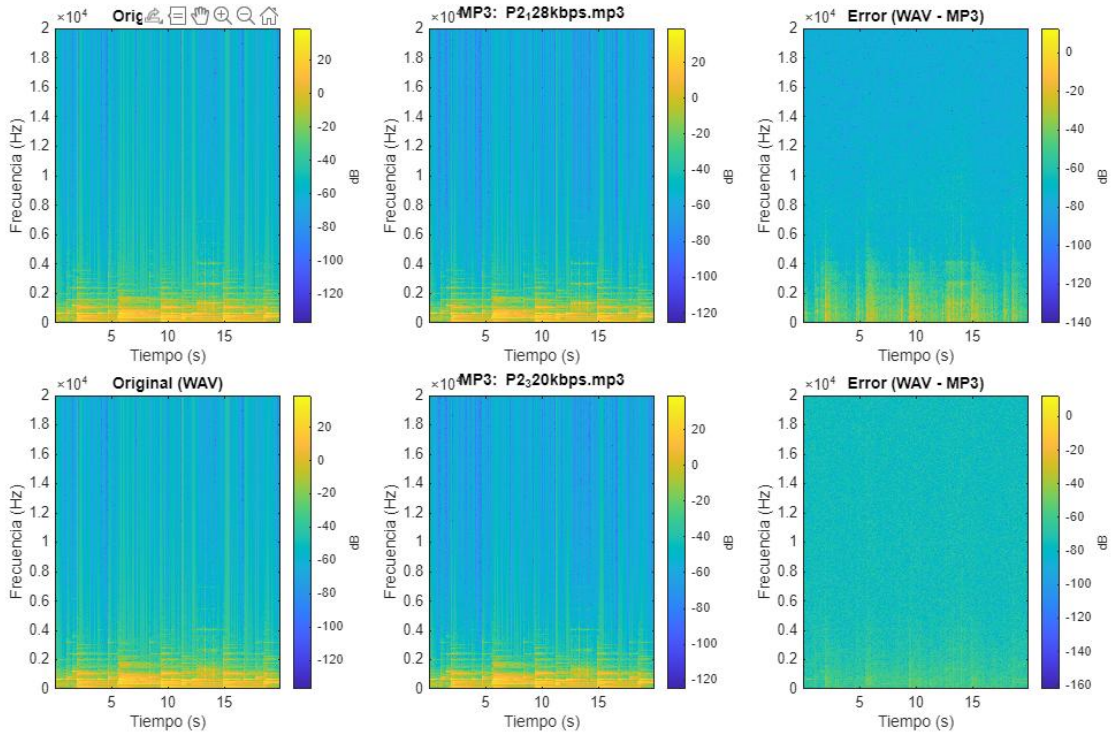

Ilustración 13: Espectrogramas MP3, WAV y Error

En 128 kbps el error (panel derecho, fila superior) muestra energía difusa en 0–3 kHz durante las colas reverberantes y una leve pérdida por encima de ~10 kHz. En 320 kbps (fila inferior) el error es más tenue y homogéneo, con mejor preservación de ataques y agudos.

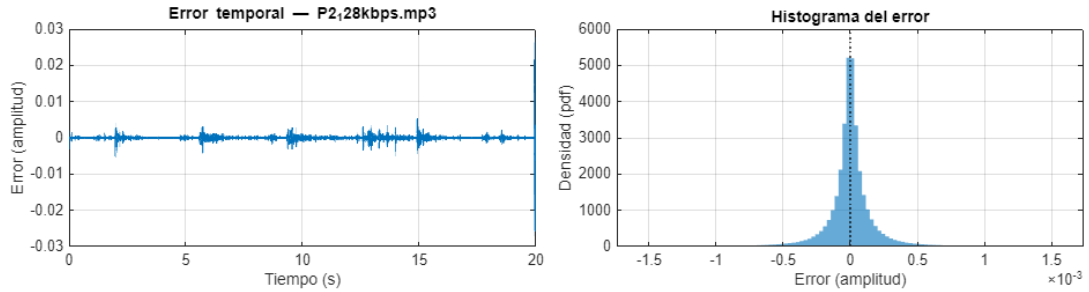

Ilustración 14: Error temporal e histograma

La señal de error es pequeña, concentrada en pasajes con sostenidos; el histograma está centrado en 0 con colas finas, lo que indica cuantización/filtrado sin sesgo. El rango del error confirma las cifras de RMSE ~6×10^{-4}.

Los dos MP3 mantienen una calidad alta para este piano reverberante. 320 kbps es consistentemente mejor (LSD ↓, SNR ↑) y conserva mejor agudos y profundidad; 128 kbps sigue siendo convincente, pero aplana ligeramente colas y brillo. En escucha casual pueden parecer equivalentes; con atención o monitores, 320 kbps resulta más fiel.

A la hora de enfocar este trabajo, se deben distinguir tres ramas de trabajo distintas: Hardware, Software y documentación.

3.4. Comentar Usar el fichero de audio en formato wav para cargarlo en 'Audacity' y desde esa aplicación exportarlo con formato mp3. Evaluar tamaños origen y final y su ratio de compresión. Cargar ambos ficheros en Matlab, tanto el original wav, como el exportado en mp3, y analizar sus diferencia o error de manera cuantitativa y de forma gráfica con gráficos e histograma.

Se cargó el archivo WAV en Audacity y se exportó en formato MP3, Posteriormente se cargaron ambos en MATLAB, se alineó el MP3 con el WAV (correlación) y se igualó el nivel RMS para medir el error.

Fichero	Tamaño (MB)	Ratio WAV/MP3	Reducción
WAV: P4audioSTM.wav	5.04	—	—
MP3 (Audacity): P4audioSTMAudacity.mp3	0.32	15.89 : 1	93.7 %

Tabla 5: Tamaños y compresión

Sobre las métricas cuantitativas entre los diferentes archivos, se tiene un error pequeño (RMSE ~$3.7 \cdot 10^{-4}$), con diferencia log-espectral moderada (LSD ≈ 2 dB). La segmental-SNR ronda 35 dB, coherente con un material estable con mucha reverberación.

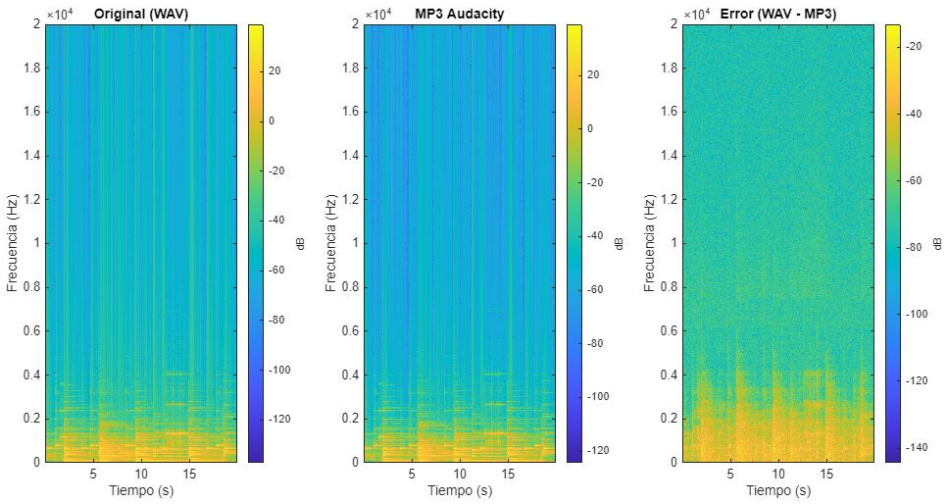

Ilustración 15: Espectrogramas

El panel de error concentra energía en graves-medios (≤3–4 kHz) durante colas reverberantes y muestra una ligera atenuación de agudos; no aparecen artefactos dominantes ni bandas espurias.

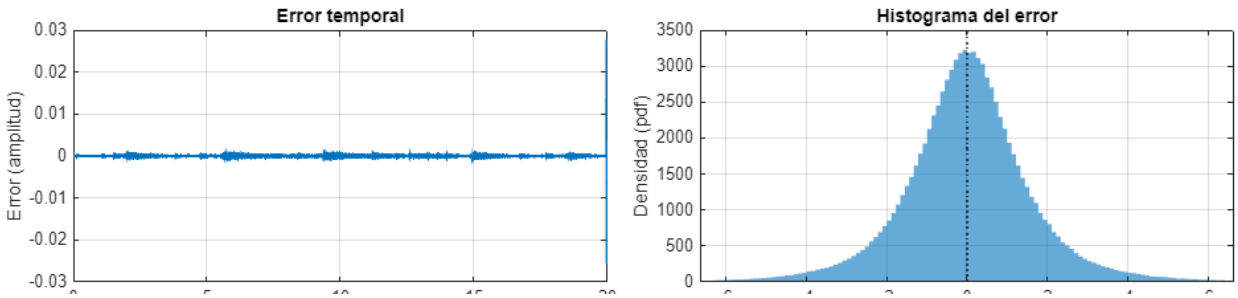

Ilustración 16: Error temporal e histograma

El error temporal, tras la alineación, es muy bajo y estable, con pequeños picos ligados a los ataques y a los sostenidos largos. El histograma está centrado en cero y presenta una distribución estrecha y simétrica, lo que indica cuantización/filtrado sin sesgo.

La exportación MP3 con Audacity logra una compresión elevada (≈ 16×) manteniendo un error bajo y sin artefactos evidentes en el espectrograma. Para este piano reverberante, el resultado es convincente: diferencias medibles pero discretas en agudos y colas, con calidad subjetiva alta.

3.5. Comentar la calidad obtenida respecto a los casos analizados antes, Evaluar las diferencias encontradas en ambos métodos de paso a mp3.

Método / Fichero	SNR (dB)	SegSNR (dB)	RMSE (–)	LSD (dB)	Lag (muestras)	Tamaño (MB)	Ratio WAV/MP3
P2_128kbps.mp3	38.60	34.95	5.75×10^{-4}	2.07	1729	0.31	16.48×
P2_320kbps.mp3	39.57	34.96	5.14×10^{-4}	1.58	1729	0.76	6.59×
Audacity MP3	42.51	34.99	3.66×10^{-4}	2.01	577	0.32	15.89×

Tabla 6: Métricas

Respecto a la tabla se puede decir que, en fidelidad espectral destaca el 320 kbps (LSD 1.58 dB), seguido muy cerca por Audacity (2.01 dB) y después 128 kbps (2.07 dB). En error global, Audacity obtiene menor RMSE y mayor SNR, con un tamaño parecido al 128 kbps. El retardo de códec difiere: 1729 muestras en LAME/FFmpeg frente a 577 en Audacity. En tamaño, 128 kbps y Audacity rondan 0.31–0.32 MB (≈16× de compresión) y 320 kbps queda en 0.76 MB (≈6.6×).

Por sacar una conclusión, si se busca máxima transparencia, el MP3 a 320 kbps es la opción más consistente (LSD mínima) aunque pese más. Para eficiencia con calidad muy alta, el MP3 de Audacity ofrece el mejor equilibrio: tamaño ~128 kbps pero métricas (SNR/RMSE) algo mejores y LSD cercana al 320 kbps. El 128 kbps de LAME/FFmpeg es el más comprimido, con pequeñas pérdidas en brillo/colas respecto a los otros dos, aunque sigue siendo sólido para escucha casual.

4. Conclusión

El análisis psicoacústico del piano con gran reverberación (Parte 1) mostró un espectro con pocos picos dominantes y una base difusa en graves-medios propia de las colas, lo que elevó los umbrales de enmascaramiento en las primeras bandas críticas y dejó márgenes (SMR) más ajustados ahí que en agudos. Los modelos de enmascaramiento y los umbrales globales en Bark fueron coherentes entre fragmentos, y la comparación por sub-bandas confirmó mínimos más relajados a medida que aumenta la frecuencia. En conjunto, la señal es "fácil" de comprimir en agudos y más exigente en la zona baja por el colchón reverberante.

En la compresión (Parte 2) se obtuvo una reducción drástica del tamaño: ≈6.6× a 320 kbps y ≈16× a 128 kbps; el MP3 exportado con Audacity quedó en ≈16× manteniendo un retardo distinto. Las métricas objetivas y las figuras respaldan lo audible: a 320 kbps se minimiza la distancia log-espectral (LSD) y se preservan mejor ataques y brillo; a 128 kbps el error aumenta ligeramente, sobre todo en colas y altas frecuencias, pero la calidad sigue siendo convincente en escucha casual. El MP3 de Audacity, con tamaño similar al 128 kbps, mostró SNR/RMSE algo mejores y una LSD muy próxima, señal de que su preset reparte bien los bits (probable VBR) pese a un retardo diferente.

Si el objetivo es máxima transparencia para material acústico reverberante, 320 kbps es la opción más sólida. Si se prioriza eficiencia con calidad alta, el MP3 de Audacity o 128 kbps ofrecen un equilibrio notable, asumiendo una ligera pérdida de "aire" y detalle en colas. En todos los casos, las diferencias quedan bien caracterizadas por las métricas (SNR, SegSNR, RMSE, LSD) y por el análisis espectro-temporal del error, que complementan la evaluación auditiva y cierran de forma consistente la práctica.

5. Anexo Código

Igual que con prácticas anteriores, se ha incluido un enlace al repositorio en Github donde se encuentra todo el contenido de la práctica (códigos, documentos y archivos de audio): https://github.com/MrAndy5/PracticasSTM/tree/main/P4