

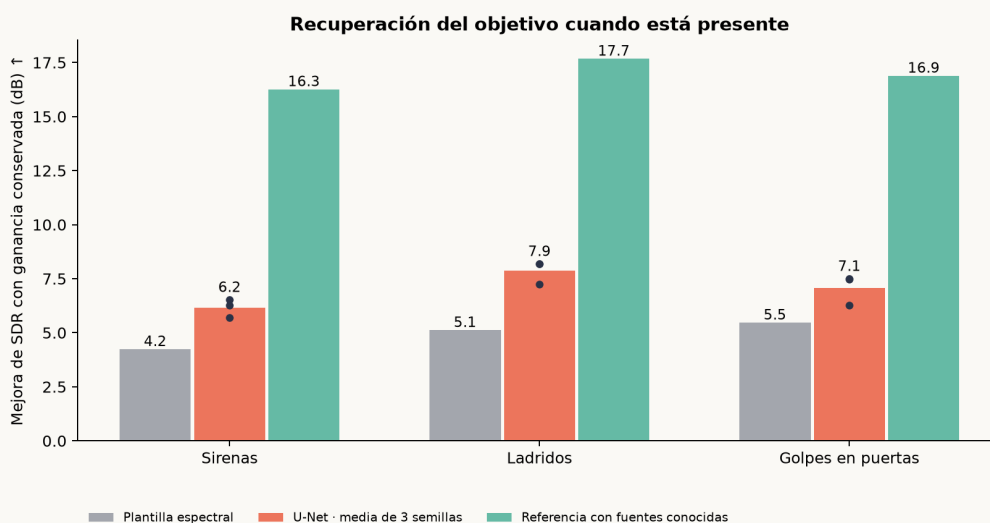
Selective Hearing

Aislar, priorizar o atenuar el sonido elegido

Una U-Net compacta, entrenada desde cero y condicionada por clase, recupera una grabación objetivo de mezclas sintéticas de sirenas, ladridos, golpes en puertas y ambiente. La misma estimación permite los tres modos de escucha.

+7.06 dB

Mejora media de SDR con ganancia conservada sobre la mezcla de entrada; media de tres semillas. La plantilla espectral obtiene 4.95 dB. Las semillas U-Net van de 6.83 a 7.43 dB.



Prueba: 300 escenas, 900 consultas; 525 con objetivo presente y 375 sin él. Los puntos marcan las tres semillas. La referencia con fuentes conocidas no es desplegable ni una cota estricta para cada métrica.

Alcance de la conclusión

FSD50K contiene grabaciones etiquetadas, no pistas de eventos puros garantizadas. La referencia incluye su fondo original. Presente/ausente se define por los componentes etiquetados de la mezcla, sin garantizar presencia o ausencia acústica en cada recorte. El resultado no demuestra aislamiento universal, calidad perceptual humana ni funcionamiento con micrófono en tiempo real.

La demo procesa archivos locales de hasta 15 segundos. Código, datos de procedencia, pesos, planes de mezclas y métricas por consulta quedan preservados.

Datos antes que mezclas

Partición	Sirenas	Ladridos	Puertas	Ambiente	Total
Entrenamiento	62	344	220	242	868
Validación	14	69	43	58	184
Prueba	55	122	87	96	360

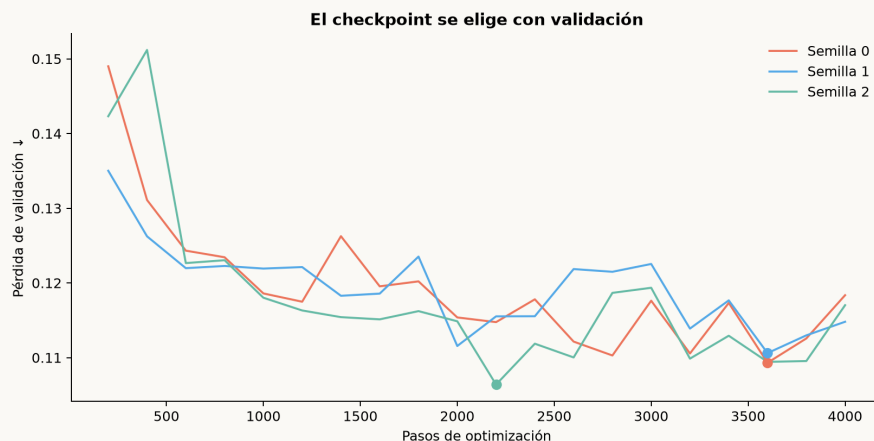
Se preservó el test oficial de FSD50K y se separó desarrollo por autor. Los tres conjuntos no comparten autores, identificadores ni contenidos exactamente iguales según hashes de original y PCM; no se descartaron casi duplicados por recodificación o recorte. Los golpes requieren una referencia a puertas en título o etiquetas. No hubo auditoría manual de escucha.

Cada escena de dos segundos contiene un ambiente y de cero a tres clases objetivo, con probabilidades 0,05 / 0,25 / 0,55 / 0,15. Se consulta cada clase en la misma mezcla. Se aplican niveles aleatorios y una escala de pico común; validación y prueba conservan planes fijos.

Red y selección

Mono 16 kHz; STFT Hann de 512 muestras, salto 128. U-Net con FiLM de clase, anchos 16/32/64/128 y 318.361 parámetros; máscara real entre 0 y 1 y fase de mezcla. Pérdida de error cuadrático complejo y de magnitud comprimidos (exponente 0,3), pesos iguales y normalización por energía comprimida de la mezcla.

AdamW, decaimiento 0,0001, recorte de gradiente 5 y lote de nueve consultas. Dos pilotos de 400 pasos seleccionaron tasa 0.001; luego tres semillas hicieron 4.000 pasos cada una. Validación: 120 escenas, 360 consultas, controles cada 200 pasos. Se retuvo la pérdida mínima, con empate a favor del paso anterior.

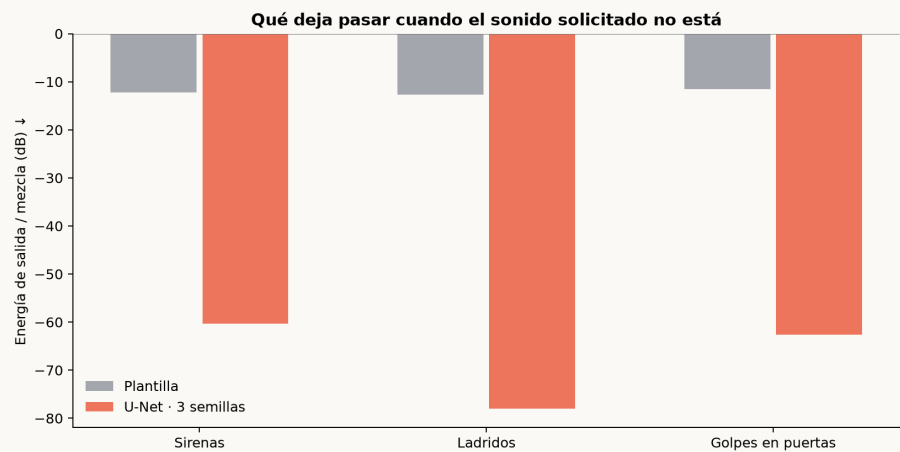


Entrenamiento y validación principales: 26.5 minutos acumulados dentro de los bucles. Excluye preparación y pilotos. La demo usa seed-0, paso 3600, por pérdida mediana de validación antes de evaluar test.

Fidelidad y fuga

Método	SDR objetivo	Mejora SDR	Mejora SI-SDR	Fuga ausente
Mezcla	-3.44	0.00	0.00	0.00
Plantilla espectral	1.51	4.95	0.83	-12.14
U-Net semilla 0	3.48	6.92	3.04	-73.01
U-Net semilla 1	3.99	7.43	3.66	-68.09
U-Net semilla 2	3.39	6.83	3.57	-59.04
U-Net media	3.62	7.06	3.42	-66.71
Fuentes conocidas	13.51	16.95	16.63	-120.00

Todas las cifras son dB. Más es mejor en SDR; una fuga más negativa es mejor ante objetivos ausentes. La plantilla se ajustó sólo con entrenamiento, dando el mismo peso a cada grabación. Las tres redes se informan por separado y promediadas; la media no es un ensamble.



Clase	Mejora SDR	Fuga media	Fuga p95
Sirenas	6.17	-60.26	-19.05
Ladridos	7.88	-78.00	-18.37
Puertas	7.08	-62.62	-15.75

Las cifras por clase son medias de los tres resúmenes por semilla. El p95 es la media de sus percentiles 95, no un intervalo de confianza. Las escenas pueden reutilizar grabaciones, por lo que 900 consultas no son 900 muestras independientes.

De la máscara a la escucha

Si x es la mezcla y s la estimación, el resultado es $a \cdot s + b \cdot (x - s)$. Aislar usa $a/b = 1/0$; priorizar, $1/0,25$; atenuar, $0/1$. Los errores de separación siguen presentes en ambos canales. Para reproducción se usa un factor común de margen; las métricas preservan las ganancias originales.

Al atenuar, el SDR del fondo conservado fue 7.06 dB. Para priorizar, el SDR de la remezcla fue 7.01 dB. Son medias entre semillas, con objetivo presente y referencias con las mismas ganancias.

El SDR del fondo coincide algebraicamente con la mejora de SDR del objetivo; no es evidencia independiente. La fuga ausente media fue -66.71 dB, pero la media de sus p95 fue -16.52 dB.

Qué significan las medidas

SDR = $10 \log_{10}(\text{energía de referencia} / \text{energía del error})$, sin centrado ni reescalado. Penaliza volumen y contenido; la salida cero puntúa 0 dB ante un objetivo presente. SI-SDR permite reescalar y se informa como complemento. Una referencia ausente no recibe SDR: se mide energía estimada / energía de mezcla. Los cocientes numéricos se limitan a ± 120 dB.

Los niveles -6/0/+6 dB fijan un objetivo ancla respecto al ambiente; otras clases también interfieren. Las fuentes cambian entre escenas, por lo que no es un barrido pareado. El cociente real objetivo/interferencia total se guarda por consulta.

Tiempo de inferencia local

Duración	Mediana CPU	Cálculo / audio
2 s	0.023 s	0.012
15 s	0.359 s	0.024

Dos hilos, un calentamiento y cinco repeticiones; excluye navegador y HTTP. El caso de 15 s repite una mezcla para medir rendimiento, no calidad en audio largo. La demo usa ventanas de 2 s con 50 % de solapamiento; la comparación principal usa escenas de 2 s.

Límites y procedencia

Tres clases fijas, sin lenguaje abierto ni separación de emisores de la misma clase. Etiquetas débiles, ambientes limitados y fase de mezcla. No se evaluaron preferencias de oyentes. La auditoría reconstruyó 1.412 PCM idénticos; 149 originales tienen muestras a escala completa y 137 remuestreados superan ± 1 , conservados en FLOAT sin recorte antes de la escala común de mezcla.

El conjunto de investigación conserva CC0, CC BY, CC BY-NC y Sampling+. Los cuatro ejemplos de validación usan sólo CC0/CC BY, con atribuciones. Se escogieron por regla fija antes de evaluar test. Las 900 consultas guardadas contienen métricas de seis métodos, no todas sus formas de onda.

Reproducir y consultar

El proyecto incluye README, GUIA_ES, protocolo, fuentes, entorno, hashes, planes, pesos y registros. El informe HTML incorpora los audios. Fuentes: [FSD50K / Zenodo](#); [artículo FSD50K](#); [SoundBeam](#); [AudioSep](#). Los dos últimos son antecedentes; no se usaron sus pesos.

Evaluación SHA-256: 21f581b209425422f06bd2f6d693569e11dc2f70a35940fe91c25e3fa7bde239