

Urdu Sukhan ASR: reconocimiento de voz preciso en urdu con la mitad del tamaño de Whisper large-v3

Asad Shehzad
Loudness.ai

Informe técnico LA-RB-2026-004, versión 1.0, 21 de octubre de 2026
doi:10.5281/zenodo.23518703

Resumen

Describimos Urdu Sukhan ASR v1, un modelo de reconocimiento de voz en urdu que también entiende inglés y la mezcla de urdu e inglés en una misma frase, y cómo se entrenó y se probó, incluidas las rondas que fallaron. El modelo es un ajuste completo de Whisper large-v3-turbo con unas 110 horas de audio transcrito por pasada: corpus públicos de urdu, dos audiolibros leídos por dos narradores, con transcripciones revisadas por el autor, clips de urdu e inglés que montamos y audio sin voz. Frente al modelo base con la misma decodificación, comete menos errores de palabra en los siete conjuntos de prueba de urdu reservados, de media un 39 % menos: del 30,9 % al 12,7 % en IndicVoices, del 39,8 % al 28,1 % con hablantes de Common Voice que nunca oyó en el entrenamiento y del 21,0 % al 16,2 % en FLEURS. En respuestas de una o dos palabras el error baja del 88,9 % al 23,6 %, en clips que pasan del urdu al inglés del 53,6 % al 19,7 %, y el inglés se mantiene en el 4,9 %. Dos de las seis rondas de entrenamiento aprendieron a oír palabras en un siseo débil; lo atribuimos a voz de entrenamiento más débil que el siseo de nuestra prueba de silencio, y mostramos que la palabra que los modelos Whisper escriben por defecto sobre el silencio procede de una regla de decodificación que prohíbe la respuesta vacía. Con esa regla desactivada, nuestro modelo no escribe nada en 100 clips de ruido, música y silencio, y ninguna puntuación con voz cambia. Damos la velocidad en cuatro máquinas con whisper.cpp, faster-whisper y Core ML, y un experimento de compresión cuyo alumno de 16 capas era 1,9 veces más rápido, pero no lo bastante preciso para publicarlo.

1. Introducción

El urdu lo hablan bastante más de 200 millones de personas, pero el reconocimiento de voz abierto para este idioma va muy por detrás del inglés. Los modelos multilingües generales como Whisper [1] vieron mucho menos urdu que inglés en su entrenamiento, y una evaluación reciente del reconocimiento en urdu [2] encontró tasas de error altas para los modelos Whisper, MMS y Seamless, sobre todo con habla conversacional. Las aplicaciones que transcriben urdu (dictado, audiolibros, entrevistas, notas de voz) se encuentran además con audio que los conjuntos de prueba públicos rara vez cubren: grabaciones bajas o de calidad telefónica, respuestas de una palabra, palabras y frases en inglés dentro del habla en urdu, y silencio, sobre el que se sabe que los modelos de tipo Whisper inventan texto [3].

Este informe describe cómo adaptamos Whisper large-v3-turbo a este contexto con un portátil y una GPU alquilada de 16 GB. El resultado conserva la velocidad y el tamaño de turbo, 809 millones de parámetros, más o menos la mitad de los 1550 millones de Whisper large-v3. Nuestras aportaciones son:

1. **Urdu Sukhan ASR v1**, en formatos whisper.cpp (q8_0, q5_0, q4_0) y CTranslate2, con una gran mejora sobre el modelo base en todos nuestros conjuntos de prueba de urdu (sección 6.1).
2. Un **ciclo de revisión** en el que cada modelo redacta transcripciones de dos audiolibros y el autor las corrige, con un estilo de la casa para el urdu aplicado a cada frase de entrenamiento y de prueba (sección 3.2).
3. La **historia completa del entrenamiento**: seis rondas, dos de las cuales fallaron con audio bajo, cinco medias de pesos descartadas y las pruebas de estrés que eligieron el punto de control publicado (secciones 4 y 6.3).
4. Un **análisis del texto inventado sobre el silencio**: tres fuentes de audio débil etiquetado en el entrenamiento y un valor por defecto de la decodificación que obliga a todo modelo Whisper a escribir al menos una palabra (sección 6.5).
5. **Mediciones de velocidad** en cuatro máquinas y tres motores, incluido el coste de la detección automática del idioma, de cada formato de cuantización y del número de hilos (sección 7).
6. Un **experimento de compresión** mediante poda del codificador y destilación, presentado como resultado parcial (sección 8), y **lecciones sobre la evaluación de modelos de voz** que cambiaron nuestras propias cifras por el camino (sección 9).

2. Antecedentes

Whisper. Whisper [1] es un transformador codificador-decodificador entrenado con 680 000 horas de audio con etiquetas débiles. El codificador lee un espectrograma log-mel de una ventana fija de 30 segundos; el decodificador escribe símbolos de texto tras un breve prefijo que indica el idioma y la tarea. Whisper large-v3-turbo conserva el codificador de 32 capas de large-v3 (128 bandas mel) y reduce el decodificador de 32 capas a 4, con 809 millones de parámetros. Su decodificador es rápido; su codificador domina el coste en una CPU.

Datos de voz en urdu. El urdu transcrito público es escaso y desigual. FLEURS [4] tiene unas siete horas de frases leídas de Wikipedia en urdu, Common Voice [5] tiene frases leídas por voluntarios de un conjunto limitado de hablantes, e IndicVoices [6] añade habla leída y conversacional de muchos hablantes de la India. El corpus Rasa de AI4Bharat tiene una parte en urdu. Varios conjuntos comunitarios en Hugging Face son obra de autores individuales.

Adaptación y destilación. LoRA [9] entrena pequeñas actualizaciones de rango bajo en lugar de todos los pesos. La media de pesos de modelos ajustados («sopas de modelos») [10] puede combinar sus puntos fuertes. La destilación del conocimiento [11] entrena un modelo pequeño para que imite a uno grande; Distil-Whisper [12] conserva todo el codificador de Whisper y reduce el decodificador. Como el decodificador de turbo ya es pequeño, un turbo más rápido en CPU tiene que quitar capas del codificador (sección 8).

Motores. whisper.cpp [13] ejecuta Whisper en CPU y en GPU de Apple; sus formatos q8_0, q5_0 y q4_0 guardan los pesos en 8, 5 y 4 bits con una escala por bloque de 32. faster-whisper [14] ejecuta Whisper sobre CTranslate2. Silero VAD [15] es un pequeño detector de actividad de voz que whisper.cpp puede ejecutar antes del modelo.

3. Datos

3.1. Fuentes y condiciones

La tabla 1 recoge las fuentes de entrenamiento del modelo publicado. Todas, salvo los dos audiolibros, son públicas. El conjunto UAT no tiene licencia pública y sus transcripciones las hizo una máquina, así que solo usamos su audio, con transcripciones hechas por un modelo nuestro anterior (sección 3.4), con el acuerdo de sus desarrolladores. La licencia propia del conjunto nzg73 permite entrenar modelos de voz citando a su autor y prohíbe usos indebidos como las suplantaciones (deepfakes). Revisamos y descartamos otros cuatro conjuntos: uno de texto a voz en urdu de Mozilla (licencia no comercial), un corpus sintético de texto a voz, una colección de podcasts sin licencia y un corpus de emociones sin transcripciones.

Fuente	Condiciones	Clips por pasada
IndicVoices urdu [6] (AI4Bharat)	CC BY 4.0	20 000
Common Voice urdu [5] (Mozilla; reempaquetado de Kaggle)	CC0	12 000
UAT, solo audio, transcrito por nuestro modelo de la ronda 4	sin licencia pública; por acuerdo	10 000
UrduSpeech (parte urdu de Rasa, de AI4Bharat)	CC BY 4.0	4999
nzg73 Urdu_Audio-Realistic (un hablante)	su propia licencia, exige citar	3000
FLEURS [4] urdu, entrenamiento y validación; inglés, entrenamiento	CC BY 4.0	2370; 2000
Dos audiolibros leídos por dos narradores; transcripciones revisadas por el autor	no se redistribuyen	1214 (521 $\times$ 2 o $\times$ 4)
Clips urdu-inglés unidos a partir de FLEURS y Common Voice	las de sus fuentes	782
Sin voz: ruido y música de MUSAN [7], siseo y zumbido generados	CC BY 4.0	2800
Total		59 165

Tabla 1: Datos de entrenamiento de Urdu Sukhan ASR v1 (ronda 6): unas 110 horas en una pasada, muestreadas de 87 571 clips depurados (unas 151 horas).

3.2. Los audiolibros y el ciclo de revisión

Se usaron dos audiolibros, cada uno leído por un narrador distinto: una traducción al urdu de un libro de autoayuda (3,97 horas) y una novela en urdu (unas 1,2 horas). Se cortaron en las pausas con Silero VAD en clips de 25 segundos como máximo (1037 clips, 305 minutos de voz) y el modelo base redactó un borrador con whisper.cpp en la GPU del portátil en 59 minutos, con una confianza para cada palabra. El autor corrigió los borradores en una herramienta web hecha para ello: las palabras dudosas aparecen en rojo o amarillo, cualquier palabra suena con un clic, las palabras corregidas automáticamente aparecen en azul y una cabecera muestra la tasa de error de los borradores en los clips ya terminados, la medida más directa de un modelo con estas grabaciones. Cada cinco clips terminados, las correcciones se revisaban y se convertían en reglas para los borradores restantes: 765 reglas al final, cada una comprobada en todos los borradores a los que afectaba, y escritas como frases completas cuando la forma errónea es en sí una palabra real.

El modelo de cada ronda volvía a redactar los clips pendientes. En el primer libro, el autor cambió el 6,1 % de las palabras de los borradores del modelo base (295 clips) y el 3,4 % de los de la ronda 1 (188 clips). En total se revisaron 579 clips (2,85 horas): 521 para entrenar y 58 reservados según su posición en la grabación (48 del primer libro y 10 del segundo); se omitieron 318 clips del primer libro y 140 del segundo conservan borradores de la ronda 3. El criterio de revisión cambió cuatro veces a petición del autor y terminó en «transcribir lo que dice el hablante»: se corrigen la ortografía, los nombres y el estilo de la casa, pero nunca se «mejora» lo que dice el hablante, para que el modelo no aprenda a corregir a nadie.

3.3. Estilo de la casa

El urdu admite varias grafías para palabras frecuentes, y entrenar con todas enseña a un modelo a ser incoherente. Aplicamos un único estilo de la casa, el del propio autor, a cada frase de entrenamiento y de prueba, y a los dos lados antes de puntuar: por ejemplo لے (liye) en lugar de لئے, چاہیے (chahiye) en lugar of چاہئے, انہیں (unhen) en lugar of انھیں, y los verbos compuestos escritos en dos palabras (kar diya, no kardiya). Se quitó la puntuación de todo el texto de entrenamiento (298 de las 299 frases de prueba de FLEURS tenían alguna), las cifras occidentales se escribieron como cifras urdu y los años dichos en voz alta como cifras. Solo esto bajó el error del modelo base en FLEURS del 22,5 % al 21,0 %: la variación ortográfica valía alrededor de 1,5 puntos.

3.4. Transcribir UAT con nuestro propio modelo

Las transcripciones de UAT coincidían palabra por palabra con la salida del modelo Whisper base en 21 de 24 clips de muestra, errores incluidos, así que las había hecho una máquina. Transcribimos sus 20 444 clips con nuestro modelo de la ronda 4 y conservamos cada clip cuyo promedio de la log-probabilidad máxima por símbolo daba una confianza de al menos 0,6, cuyo texto no era repetitivo y cuya longitud encajaba con el audio: 20 235 clips. La ronda 5 usó los de confianza $\geq 0,8$ y la ronda 6 los de confianza $\geq 0,85$ (19 721).

3.5. Depuración

Antes de la ronda 6 se eliminó todo clip de menos de 0,5 o más de 30 segundos, sin texto, con más de 22 letras por segundo (25 para clips de menos de 1,5 segundos) o menos de 1,5 por segundo si duraba más de 4 segundos (audio que no puede contener su texto), o en urdu con menos de un 60 % del texto en escritura urdu. Así se eliminaron 27 clips de Common Voice cuyo audio terminaba antes que su texto, 81 clips de IndicVoices en otra escritura, 488 clips de UAT con transcripciones dudosas y unas pocas decenas más. Los clips sin voz y de mezcla de idiomas quedaron exentos.

3.6. Clips sin voz y de mezcla de idiomas

La ronda 6 añadió 2800 clips con la transcripción vacía: 2500 fragmentos de ruido y música instrumental de MUSAN y 300 clips generados de silencio digital, siseo de micrófono (de -80 a -45 dBFS) y zumbido de la red eléctrica. También añadió 782 clips que unen un clip en urdu (FLEURS o Common Voice) y un clip en inglés de FLEURS con una breve pausa, con el urdu primero en el 70 % de ellos, transcritos con la parte inglesa en letras latinas bajo la etiqueta de idioma urdu.

3.7. Conjuntos de prueba y de estrés

Siete conjuntos de prueba nunca se usaron para entrenar: la validación de IndicVoices en urdu (300 clips, como máximo tres por hablante), clips de Common Voice en urdu de 15 hablantes ausentes de nuestros datos de entrenamiento (260), la prueba de FLEURS en urdu (299), la partición de prueba de UrduSpeech (100), clips reservados de nzs73 (300; el mismo y único hablante que su parte de entrenamiento) y capítulos reservados de los dos audiolibros (48 y 10 clips). Nueve pruebas de estrés miden la robustez: ruido (40) y música instrumental (30) de archivos de MUSAN que reservamos (todos aquellos cuyo nombre tiene un MD5 divisible por diez); silencio, siseo y zumbido (30); voz baja, clips de prueba bajados 30 dB (40); los mismos clips a través de una línea telefónica (40); respuestas de una o dos palabras como جی (ji) y ہاں (haan) (40); inglés (30); un clip de prueba en urdu seguido de otro en inglés (20); y de 25 a 30 segundos de voz de prueba unida (15).

4. Entrenamiento

4.1. Ajustes comunes a todas las rondas

Todas las rondas se entrenaron con AdamW (decaimiento de pesos 0,01; la versión de 8 bits de bitsandbytes en los ajustes completos), un 10 % de calentamiento lineal y descenso lineal hasta cero, recorte del gradiente en 1,0 y semilla 0, con etiquetas `<|ur|><|transcribe|><|notimestamps|>` texto `<|endoftext|>`. Las ejecuciones con LoRA usaron rango 32, alfa 64 y abandono de 0,05 en las proyecciones de atención y las capas de propagación, en bf16; los ajustes completos mantuvieron pesos maestros en fp32 con cálculo en bf16 y puntos de control del gradiente. Desde la ronda 3, el 30 % de los clips con tiempos de palabra conocidos se entrenaron con marcas de tiempo, para que la decodificación con marcas de tiempo siguiera funcionando.

4.2. Seis rondas

La tabla 2 resume las rondas. Las rondas 1 a 3 se hicieron en el portátil (Apple M4, 16 GB) con LoRA solo en el decodificador, con las salidas del codificador calculadas una vez y guardadas, como hace Distil-Whisper; enseñaron al modelo las dos voces de los audiolibros. Las rondas 4 a 6 usaron una RTX A4000 alquilada (16 GB). La ronda 4 entrenó LoRA en el codificador y el decodificador con aumento de audio y añadió nzs73 y todo FLEURS. La ronda 5 fue el primer ajuste completo de los 809 millones de pesos y añadió Common Voice y el audio de UAT transcrito. La ronda 6, el modelo publicado, añadió IndicVoices, UrduSpeech, el inglés de FLEURS, los clips de mezcla de idiomas y sin voz y el ruido de fondo de MUSAN, usó la mitad de la tasa de aprendizaje de la ronda 5 y completó una sola pasada con un punto de control cada 925 pasos.

4.3. Aumento de datos

Desde la ronda 4 se aumentó la mitad de los clips: velocidad de 0,9 a 1,1, un cambio de ganancia, eco de sala (una respuesta al impulso de ruido con caída exponencial) en el 30 % y ruido de fondo en el 60 %, entre 5 y 30 dB por debajo de la voz. Las rondas 5 y 6 añadieron SpecAugment [8] (máscaras de tiempo y frecuencia del 5 %). La ronda 6 cambió tres cosas: uno de cada cinco clips aumentados se hizo «muy bajo» (entre 12 y 30 dB menos), el 15 % pasó por una línea telefónica (8 kHz con saturación) y el 70 % del ruido de fondo procedía de 1590 grabaciones de MUSAN. La sección 6.5 muestra que dos de estos ajustes fueron errores.

Ronda	Método	Datos por pasada	Pasos $\times$ lote	TA	Tiempo
R1	LoRA, decodificador (portátil)	218 clips de libro $\times 4$, 600 de FLEURS	368×8	2e-4	40 min
R2	LoRA, decodificador (portátil)	365 clips de libro $\times 3$, 600 de FLEURS	422×8	2e-4	73 min
R3	LoRA, decodificador, marcas de tiempo (portátil)	435 + 85 clips de libro ($\times 2$, $\times 5$), 600 de FLEURS	472×8	2e-4	67 min
R4	LoRA, codificador y decodificador	libros, nsg73 3000, FLEURS 2109; 6409 clips	800×16	1e-4	99 min
R5	ajuste completo	datos de R4, Common Voice 8000, UAT 3000; 17 409	2176×16	1e-5	164 min
R6	ajuste completo	tabla 1; 59 165 clips	3697×16	5e-6	268 min

Tabla 2: Rondas de entrenamiento (TA: tasa de aprendizaje). R1–R3 se hicieron en la GPU del portátil y R4–R6 en una RTX A4000. La ronda 6 entrenó durante 4,5 horas de GPU (4,4s por paso).

4.4. Elegir el punto de control

Antes de la ronda 6 fijamos un umbral de aprobación: sin bucles, texto inventado como mucho en unos pocos clips sin voz, mejor que la media R4+R5 en las pruebas de estrés de voz baja, teléfono, respuestas cortas, mezcla y voz larga, y al menos igual en los siete conjuntos de prueba. Cada punto de control se puntuó en los dieciséis conjuntos; las pruebas de estrés descartaron puntos de control que parecían buenos en los conjuntos de prueba (sección 6.4).

5. Montaje experimental

Equipos. Portátil: un Apple MacBook Air con chip M4 y 16 GB de memoria, usado para los borradores, la revisión, las rondas 1 a 3 y las pruebas de whisper.cpp con Metal y Core ML. Servidor de GPU (alquilado): una NVIDIA RTX A4000 de 16 GB, cuya CPU es un Intel Xeon E3-12xx v2 de 2012 (Ivy Bridge, con AVX pero sin AVX2). Servidor en la nube: 12 núcleos virtuales AMD EPYC a 2,0 GHz (AVX2) con otros servicios en marcha, el tipo de máquina que usaría un producto.

Software. PyTorch 2.14.1 con Transformers y bitsandbytes 0.50.2; whisper.cpp v1.9.5 (versiones para CPU, Metal y Core ML) con Silero VAD v5.1.2; CTranslate2 4.5.0 para la conversión y faster-whisper 1.2.1; coremltools 9.0; Chartist 2.1.0 [16] para las figuras.

Métricas. Damos las palabras erróneas (tasa de error de palabras, WER) y las letras erróneas (tasa de error de caracteres, CER, sin contar los espacios) tras la misma normalización de la referencia y de la salida: se unifican la puntuación, los diacríticos, las formas de letra árabes y urdu, los sistemas de cifras y las variantes del hamza, y se aplica el estilo de la casa a los dos lados. Un *bucle* es una salida de más del doble de longitud que la referencia (y de al menos siete

palabras), o una frase de una a seis palabras repetida cuatro o más veces seguidas más veces de lo que la dijo el hablante; las tablas calculadas con una regla anterior, más laxa, cuentan de más los bucles con voz (11 en lugar de 1 en IndicVoices para el modelo publicado), así que solo damos bucles en clips sin voz y donde se volvieron a contar. *Texto inventado* son dos o más palabras en un clip sin voz; *cualquier palabra* cuenta los clips sin voz con al menos una palabra. En el capítulo de prueba del segundo audiolibro contamos además cuántas de sus 23 apariciones de nombres se escriben exactamente bien.

Decodificación. Las evaluaciones en GPU usan `generate` de Transformers con decodificación voraz, el idioma fijado en urdu (`auto` para las pruebas de estrés en inglés y de mezcla) y sin marcas de tiempo. Las evaluaciones con `whisper.cpp` se ejecutan como lo haría una aplicación: búsqueda en haz de anchura 5 con reintento con temperatura, sin marcas de tiempo, sin arrastrar contexto, con detección automática del idioma y el formato `q8_0`, en tres variantes: ajustes por defecto, el detector de voz Silero antes (umbral 0,5, silencio mínimo de 2s, margen de 300 ms) y la supresión de la respuesta vacía desactivada (sección 6.5).

6. Resultados

6.1. Frente al modelo base

La tabla 3 y la figura 1 comparan el modelo base y el nuestro con una decodificación idéntica. Nuestro modelo es mejor en todos los conjuntos de urdu. Las mayores mejoras relativas están en el habla conversacional de muchos hablantes (IndicVoices, un 59 % menos de errores de palabra), en el único hablante de `nzg73` (67 %), en las respuestas cortas (73 %) y en los clips de urdu e inglés (63 %). FLEURS, texto leído de Wikipedia con muchos nombres y palabras técnicas, es el conjunto público que menos mejora (23 %), igual que el segundo audiolibro (23 %), cuyo narrador lee muchos nombres (sección 6.6). El inglés no cambia y se queda en el 4,9 %.

6.2. Ejemplos

Las tasas de error no muestran cómo son los errores. Pasamos 47 clips de prueba públicos con voz, y dos sin voz, por los dos modelos como lo haría una aplicación (`whisper.cpp`, `q8_0`, haz de 5, reintento con temperatura). En estos clips, nuestro modelo fue mejor en 27, peor en 4 e igual en 16 (19,1 % frente a 37,8 % de palabras erróneas). La tabla 4 muestra clips cortos que elegimos a mano para enseñar dónde más ayuda el ajuste, de lo alto de una clasificación de todos los clips por palabras corregidas, y uno de los 4 clips en los que salió peor; la ficha del modelo los tiene con su audio. El modelo base convierte las respuestas cortas en otras palabras, incluso en «música» por un «hola» hablado y en «Alá, Alá» por «por separado». Las mayores mejoras fueron en clips que pasan del urdu al inglés, demasiado largos para la tabla: en uno, el modelo base escribió cuatro palabras en inglés para un clip de 28 palabras; en otro, tradujo al inglés la mitad en urdu en lugar de transcribirla. El nuestro transcribió los dos, con dos palabras erróneas en total. Aún falla en la separación de algunas palabras, como muestra el último ejemplo. Sobre el silencio, permitir un resultado vacío solo ayuda a nuestro modelo: el modelo base sigue escribiendo una palabra.

6.3. Las rondas de entrenamiento

Rondas 1 a 3: las voces de los audiolibros. Las tres primeras rondas entrenaron con unos cientos de clips de los libros y 600 clips de FLEURS. Bajaron el error en los libros (tabla 5)

Conjunto (clips)	WER (%)		CER (%)	
	Base	Nuestro	Base	Nuestro
IndicVoices urdu, validación (300)	30,9	12,7	21,9	7,5
Common Voice, hablantes no vistos (260)	39,8	28,1	25,3	16,4
FLEURS urdu, prueba (299)	21,0	16,2	8,3	6,3
UrduSpeech, partición de prueba (100)	14,5	8,8	5,8	2,1
nzg73, clips reservados (300)	4,8	1,6	1,7	0,6
Audiolibro 1, capítulos reservados (48)	5,9	4,0	2,6	1,6
Audiolibro 2, capítulo reservado (10)	17,2	13,2	6,0	4,8
Voz baja (40)	35,6	23,5	24,8	15,9
Voz de banda telefónica (40)	36,2	25,0	21,8	14,8
Respuestas de una o dos palabras (40)	88,9	23,6	66,7	27,2
Urdu y luego inglés en un clip (20)	53,6	19,7	49,8	14,2
Voz continua, 25–30 s (15)	35,0	29,8	26,0	20,7
Inglés, FLEURS (30)	4,9	4,9	2,0	1,7
Texto inventado: ruido / música / silencio	15 / 16 / 1	6 / 4 / 3		

Tabla 3: Modelo base (Whisper large-v3-turbo) y Urdu Sukhan ASR v1 en los conjuntos de prueba reservados (arriba) y en las pruebas de estrés (abajo); Transformers, decodificación voraz; cuanto menos, mejor; el mejor valor, en negrita. La última fila cuenta los clips sin voz con dos o más palabras.

pero no con otras voces: con hablantes de Common Voice, la ronda 3 obtuvo un 37,0 % frente al 37,1 % del modelo base (con el archivo de prueba anterior, que tenía filas duplicadas). La ronda 2 mostró además que un modelo entrenado sin marcas de tiempo y decodificado con ellas (el valor por defecto de whisper.cpp) puede detenerse en una pausa larga y perder el resto de un clip: 3 de 40 clips de prueba se cortaron, uno de ellos en 59 palabras. Decodificar sin marcas de tiempo lo resolvió (del 4,8 % al 3,2 %), y desde la ronda 3 el 30 % de los clips de entrenamiento llevaba marcas de tiempo. Estos primeros conjuntos de prueba cambiaron de una ronda a otra y se revisaron a partir de borradores de la ronda 1, lo que favorece a la ronda 1; no son comparables con las tablas posteriores.

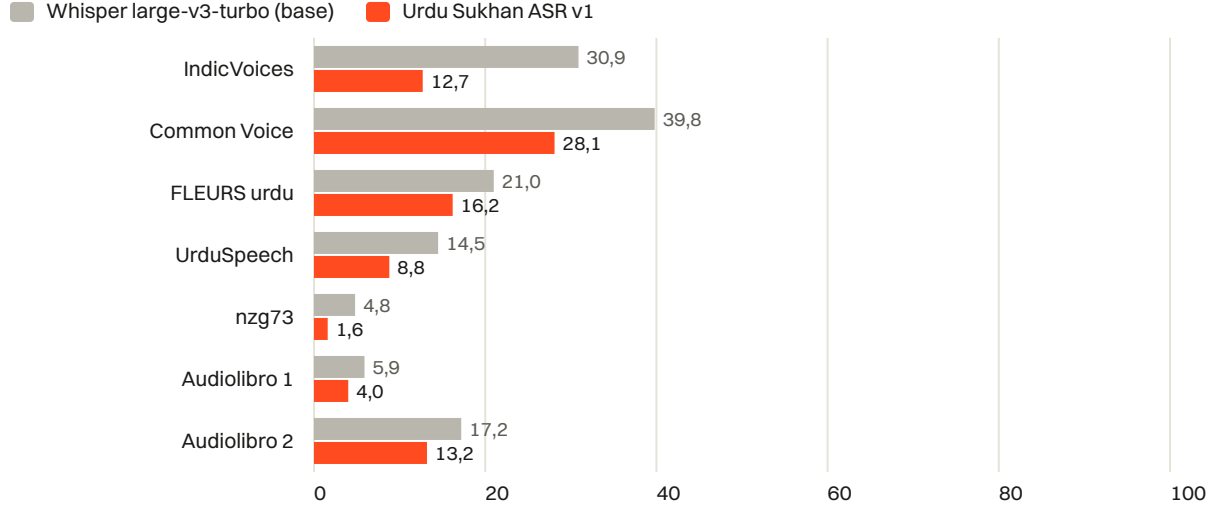
Rondas 4 y 5: urdu general y luego un fallo. La ronda 4 fue la primera que mejoró el urdu general: FLEURS del 21,0 % al 17,6 % y los hablantes no vistos de Common Voice del 39,8 % al 31,7 % (tabla 6). La ronda 5, el primer ajuste completo, fue el modelo más preciso hasta entonces con habla leída, pero en los clips no vistos de Common Voice escribió repeticiones descontroladas en 21 de 260 clips, 17 de ellos de un mismo hablante de voz baja con clips muy cortos, y obtuvo un 149,3 % de palabras erróneas. La precisión numérica y el volumen se descartaron volviendo a puntuar en fp16 y tras normalizar los picos. En las pruebas de estrés inventó texto en 30 de 30 clips de silencio y en 38 de 40 de ruido (figura 2). Las causas probables eran la falta de datos de entrenamiento sin voz, clips de Common Voice cuyo audio termina antes que su texto y un ajuste completo que se sobreajustaba al audio inusual; no las probamos una a una. Con reintento con temperatura en whisper.cpp, la ronda 5 no entró en bucle y obtuvo un 27,1 % en los mismos clips, lo que muestra cuánto ocultan los ajustes del decodificador.

La media R4+R5. Una media al 50 % de los pesos de las rondas 4 y 5 conservó la mayor parte de las mejoras de la ronda 5 y entró en bucle solo en 3 de 260 clips de Common Voice (una media

Clip y significado	Sistema	Transcripción
Respuesta de dos palabras, IndicVoices	Referencia	الگ الگ
«por separado»	large-v3-turbo	اللہ اللہ
	Urdu Sukhan	الگ الگ
Respuesta de una palabra, IndicVoices	Referencia	ہیلو
«hola»	large-v3-turbo	موسیقی
	Urdu Sukhan	ہیلو
IndicVoices	Referencia	جی جی جی ہاں
«sí, sí, sí, sí»	large-v3-turbo	دی دی دی ہاں
	Urdu Sukhan	جی جی جی ہاں
Voz baja	Referencia	ٹام کار کے حادثے میں بچ گیا تھا
«Tom sobrevivió al accidente de coche»	large-v3-turbo	ٹوم کار کی ہاتھ سے میں بچ گیا تھا
	Urdu Sukhan	ٹوم کار کے حادثے میں بچ گیا تھا
Línea telefónica	Referencia	اس ماحول میں جو آزادی نظر آتی ہے
«la libertad que se ve en este ambiente»	large-v3-turbo	اس محول میں جو آزادی نظر آتی ہے
	Urdu Sukhan	اس ماحول میں جو آزادی نظر آتی ہے
UrduSpeech, el nuestro peor	Referencia	گیا آند وہار غریب رتھ ایکسپریس
«Gaya–Anand Vihar Garib Rath Express (un tren)»	large-v3-turbo	گیا آند وہار غریب رتھ ایکسپریس
	Urdu Sukhan	گیا آند وہار [-] غریب رتھ ایکسپریس
Siseo débil, sin voz	large-v3-turbo	موسیقی
	Urdu Sukhan	اچھا
	Urdu Sukhan, vacía permitida	(nada)

Tabla 4: Ejemplos de conjuntos de prueba públicos (whisper.cpp, q8_0, haz de 5). Las palabras erróneas están en rojo y una palabra que falta aparece como [-]; la grafía y la separación de palabras se comparan con la referencia del conjunto de prueba.

Conjuntos de prueba reservados



Pruebas de estrés

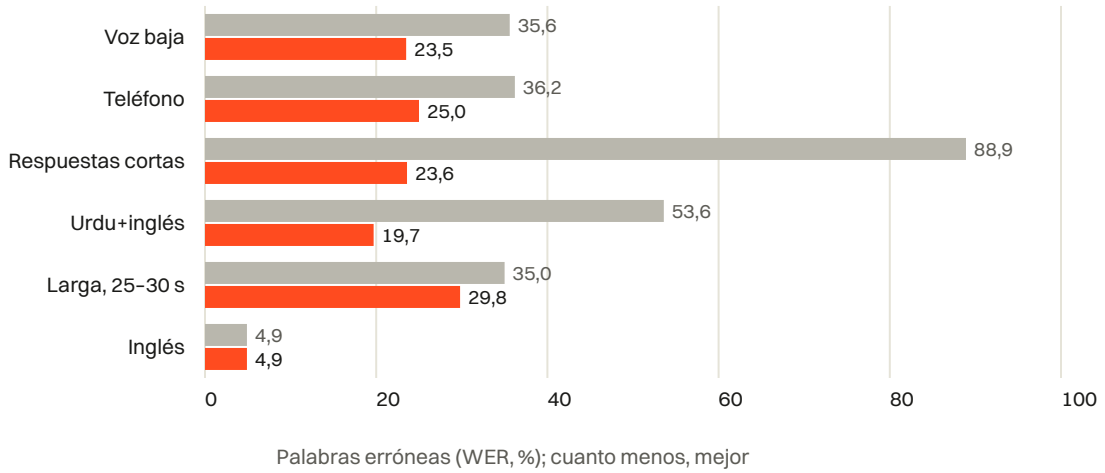


Figura 1: Palabras erróneas del modelo base (gris) y de Urdu Sukhan ASR v1 (naranja) en cada conjunto de prueba y de estrés con voz.

WER (%)	Base	R1	R2	R3	R4
Libro 1, 40 clips (rondas 1-2)	5,5	2,7	3,2	—	—
Libro 1, 48 clips	5,9	3,9	3,9	4,3	3,8
Libro 2, 10 clips	17,0	15,1	15,7	13,2	11,7

Tabla 5: Rondas 1 a 4 en los clips de prueba de los audiolibros (whisper.cpp q8_0, haz de 5, sin marcas de tiempo, GPU del portátil).

30/70 lo hizo en 9). Fue nuestro mejor modelo hasta la ronda 6 y sigue siendo el mejor en el segundo audiolibro (9,7 % frente a 13,2 %), pero inventó texto en 35 de 100 clips sin voz y no superó las pruebas de mezcla y de respuestas cortas.

Ronda 6. La ronda 6 entrenó con cinco veces más datos, con clips sin voz y de mezcla de idiomas, y en el paso 2775 es la mejor o casi la mejor en todos los conjuntos. Es la primera ronda que escribe las palabras inglesas en letras latinas (sección 6.7).

WER (%) / clips	Base	R4	R5	R4+R5	R6/925	R6/1850	R6/2775	R6/3697
IndicVoices	30,9	19,6	19,7	22,3	14,5	15,0	12,7	13,0
Common Voice	39,8	31,7	149,3	56,0	38,9	37,9	28,1	92,6
FLEURS urdu	21,0	17,6	15,8	15,7	20,9	17,2	16,2	15,4
UrduSpeech	14,5	12,2	12,0	11,2	11,8	9,8	8,8	8,8
Audiolibro 1	5,9	4,1	3,5	3,7	5,6	5,4	4,0	3,9
Audiolibro 2	17,2	11,6	11,1	9,7	17,4	13,6	13,2	13,5
Voz baja	35,6	27,8	93,5	26,1	28,1	24,5	23,5	271,6
Teléfono	36,2	29,3	169,0	26,1	27,3	23,3	25,0	135,3
Respuestas cortas	88,9	69,4	73,6	65,3	31,9	25,0	23,6	26,4
Urdu+inglés	53,6	65,6	103,4	72,8	22,7	45,8	19,7	19,7
Larga (25–30 s)	35,0	29,6	29,6	27,9	33,9	30,4	29,8	29,3
Inglés	4,9	4,8	5,6	5,1	5,2	5,1	4,9	4,9
Ruido: inventado	15	12	38	14	2	8	6	21
Música: inventado	16	14	8	10	1	2	4	2
Silencio: inventado	1	9	30	11	0	0	3	27
Bucles en clips sin voz	3	3	35	4	3	5	5	36

Tabla 6: Cada ronda en los conjuntos de prueba y de estrés comunes (WER %, Transformers, voraz). R4: LoRA; R5: ajuste completo; R4+R5: media de pesos de las dos; R6/ n : ronda 6 tras n pasos. El mejor valor de cada fila, en negrita. nsg73 se excluye porque las rondas 4 y 5 usaron una parte reservada más pequeña. Bucles contados con la primera regla.

6.4. Elegir el punto de control

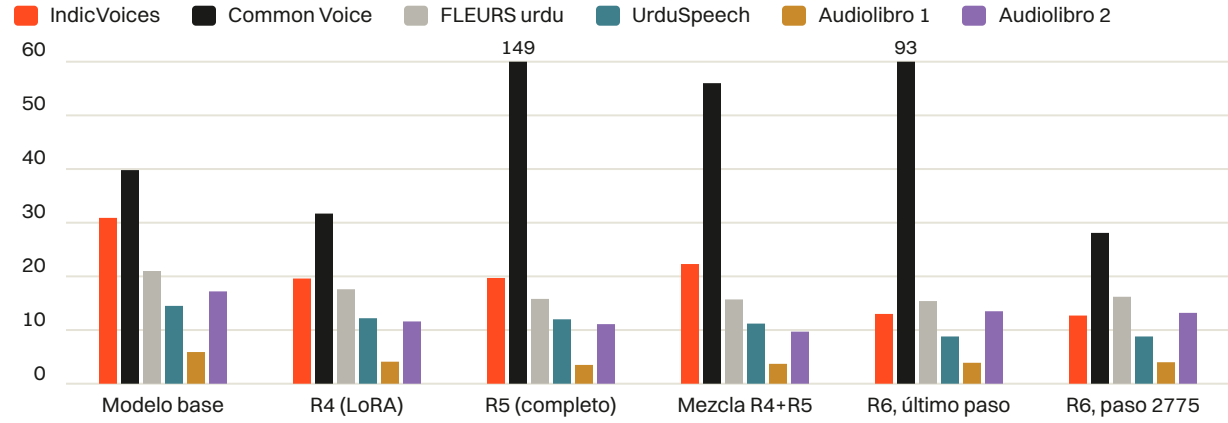
La precisión con voz mejoró hasta el paso 2775 (figura 3). En los últimos 922 pasos, mientras la tasa de aprendizaje bajaba a cero, la ronda 6 aprendió a oír palabras en el audio débil: el texto inventado sobre el silencio subió de 3 a 27 de 30 clips y sobre el ruido de 6 a 21 de 40, la voz baja pasó del 23,5 % al 271,6 % de palabras erróneas y Common Voice del 28,1 % al 92,6 % con 7 bucles. Publicamos el paso 2775. Los puntos de control anteriores eran menos precisos: el paso 925 obtuvo un 20,9 % en FLEURS y el paso 1850 un 45,8 % en el conjunto de urdu e inglés.

La media de pesos no ayudó. Promediamos el paso 2775 con el último paso, y con la media R4+R5 al 25 % y al 50 % (tabla 7). Cada media ganó hasta 2,5 puntos en el segundo audiolibro y perdió en el resto: las medias inventaron texto en 12 a 17 clips de silencio y duplicaron o triplicaron los errores con voz telefónica. No se conservó ninguna.

6.5. Texto inventado sobre el silencio

Audio débil etiquetado. La ronda 5 y el final de la ronda 6 aprendieron que un siseo débil significa voz. Medir el nivel del audio explica por qué (figura 4). Las grabaciones de FLEURS en

Palabras erróneas en los conjuntos de prueba reservados (WER, %)



Las barras se cortan en el 60 %; el valor de una barra cortada aparece encima (R5 (completo): Common Voice 149,3 %; R6, último paso: Common Voice 92,6 %).

Clips de estrés sin voz con texto inventado (dos o más palabras), de 100

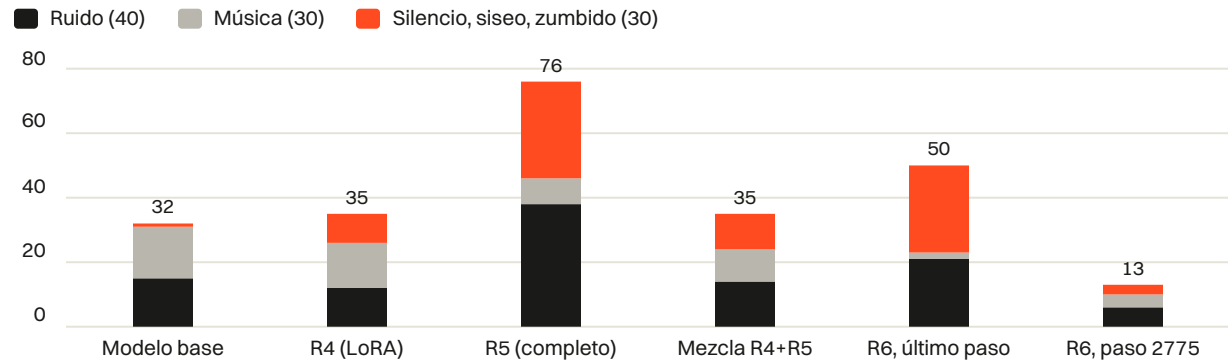


Figura 2: Palabras erróneas en seis conjuntos de prueba por ronda (arriba) y clips de estrés sin voz en los que cada ronda escribió dos o más palabras (abajo).

WER (%) / clips	R6/2775	tardía	+25 % mezcla	+50 % mezcla
Common Voice	28,1	55,1	34,1	130,3
IndicVoices	12,7	13,5	12,0	13,2
Audiolibro 2	13,2	12,3	11,4	10,7
Voz baja	23,5	149,7	24,2	83,7
Teléfono	25,0	80,2	61,8	79,9
Urdu+inglés	19,7	19,7	62,2	70,2
Ruido: inventado	6	17	20	28
Música: inventado	4	4	1	1
Silencio: inventado	3	12	17	14

Tabla 7: Medias de pesos de la ronda 6 (WER % y clips con texto inventado; Transformers, voraz). «Tardía»: el paso 2775 y el último paso; «mezcla»: el paso 2775 con la media R4+R5.

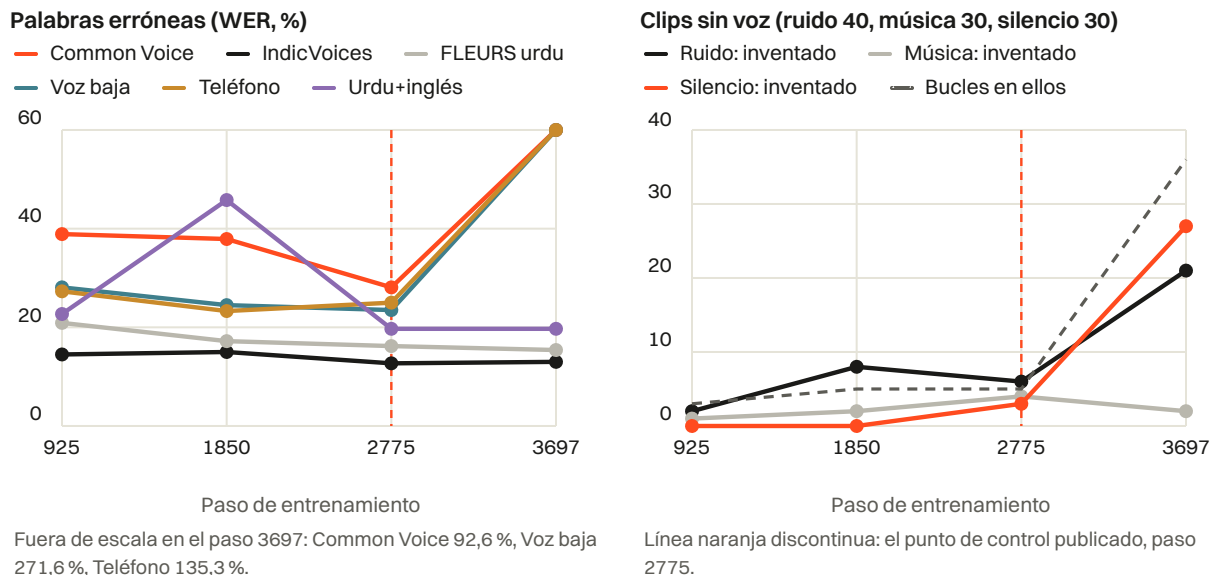


Figura 3: Puntos de control de la ronda 6: palabras erróneas en tres conjuntos de prueba y tres de estrés (izquierda) y clips sin voz con texto inventado, con sus bucles (derecha).

urdu son muy bajas: mediana de -45 dBFS, con uno de cada veinte clips por debajo de -60 dBFS, lo que se solapa con el siseo de nuestra prueba de silencio (-66 a -52 dBFS). El aumento de datos de la ronda 6 añadió dos fuentes más. El ajuste «muy bajo» redujo una quinta parte de los clips aumentados entre 12 y 30 dB, con ruido de fondo de hasta el mismo volumen que la voz. Y el filtro de eco de sala normalizaba su respuesta al impulso por su suma, lo que baja el nivel de un clip con eco otros 20 a 25 dB. Juntos dejaron voz *con transcripción completa* entre -40 y -75 dBFS, por debajo del siseo de la prueba de silencio, así que un modelo premiado por transcribirla aprende a transcribir el siseo. La solución es llevar primero cada clip a un nivel normal, aplicar el nivel final después del filtro de eco, no bajar nunca de -40 dBFS y mantener el ruido entre 15 y 35 dB por debajo de la voz baja. La usamos en el experimento de compresión (sección 8) y la usaremos en las próximas rondas.

La respuesta de una palabra es una regla de decodificación. Por defecto, whisper.cpp (`suppress_blank`) y Transformers (`begin_suppress_tokens`) prohíben el símbolo de fin de texto y un espacio como *primer* símbolo generado, así que un modelo Whisper no puede devolver una transcripción vacía. Medido justo después del prefijo en clips sin voz, nuestro modelo pone casi toda su probabilidad en un espacio, porque el código de entrenamiento codificaba la transcripción vacía como un único símbolo de espacio; el modelo comprimido de la sección 8 pone una probabilidad de 1,00 en el fin de texto. Ambas respuestas son justo los símbolos prohibidos, así que por defecto el modelo tiene que escribir una palabra, normalmente *اچھا* (*achha*, «de acuerdo») o *ہاں* (*haan*, «sí»), y con música *موسیقی* (*mosiqi*, «música»). La medida habitual de dos palabras lo oculta: por defecto, tanto el modelo base como el nuestro escriben al menos una palabra en cada clip de ruido, música y silencio (figura 5). Con la supresión desactivada (una línea en una aplicación: `params.suppress_blank = false`), nuestro modelo no escribe nada en los 100 clips y ninguna puntuación con voz cambia (tabla 8).

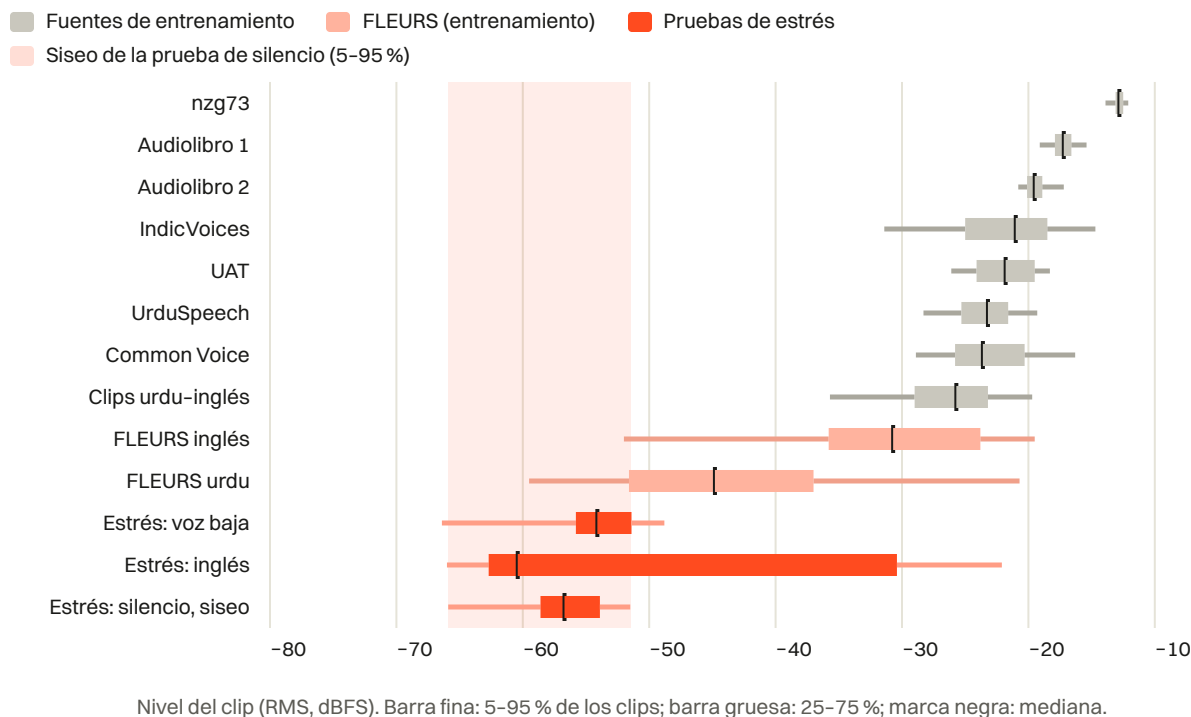


Figura 4: Niveles de los clips (RMS) de hasta 300 clips al azar por fuente de entrenamiento y de tres pruebas de estrés. FLEURS es la fuente de entrenamiento más débil; la franja marca el siseo de la prueba de silencio. Se omiten 10 clips de silencio digital puro. Los clips de estrés en inglés también son débiles, lo que importa para el detector de voz.

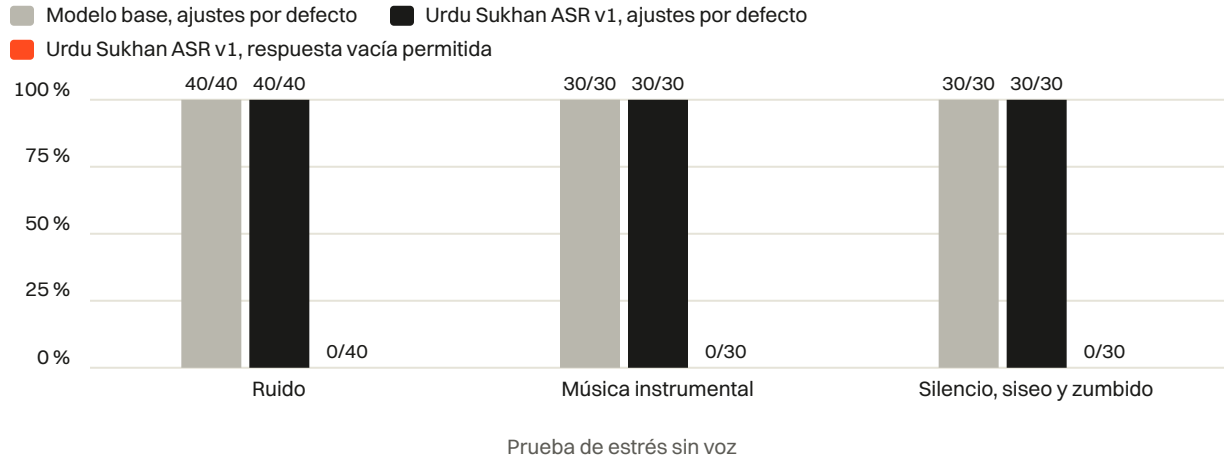
El detector de voz. Pasar antes Silero VAD también elimina casi todo el texto inventado, pero baja el resultado en el conjunto de urdu e inglés del 20,4 % al 44,6 %. Las mitades inglesas de ese conjunto son grabaciones de FLEURS en inglés con su nivel natural (una a -69 dBFS frente a -23 dBFS de su mitad en urdu), y el detector las descarta como silencio: un defecto del conjunto de prueba más que del detector.

6.6. Nombres

Los nombres son el punto más débil. El capítulo de prueba del segundo audiolibro tiene 23 apariciones de nombres; el modelo base escribió bien 4 y el nuestro 11 con decodificación voraz (la media $R4+R5$, 15). Sus 86 clips de entrenamiento contienen los nombres *عمير* (*Umais*) y *عزير* (*Uzair*) una vez cada uno, y el modelo los escribe como la palabra común *امير* (*ameer*, «rico») y como *اوزير*; también escribe *دانيا* por *تانيا* (*Tania*). Una indicación de whisper.cpp con la lista de nombres del capítulo subió la cuenta de 12 a 17 de 23 sin bucles (tabla 9).

6.7. El inglés dentro del urdu

Las rondas 1 a 5 se entrenaron solo con transcripciones en escritura urdu y escribían las palabras inglesas en letras urdu. La ronda 6 fue la primera entrenada con clips de urdu e inglés cuya parte inglesa se escribe en letras latinas bajo la etiqueta de idioma urdu, y con 2000 clips de FLEURS en inglés, así que mantiene las palabras inglesas en letras latinas dentro del texto en urdu. El autor lo



Proporción de clips sin voz en los que whisper.cpp escribió al menos una palabra (q8_0, búsqueda en haz de 5, reintento con temperatura). «Respuesta vacía permitida»: suppress_blank desactivado.

Figura 5: Clips sin voz en los que whisper.cpp escribió al menos una palabra: el modelo base y el nuestro con los ajustes por defecto, y el nuestro sin la supresión de la respuesta vacía.

Prueba de estrés (clips)	large-v3-turbo	Urdu Sukhan ASR v1		
	por defecto	por defecto	detector de voz	vacía permitida
Ruido (40)	16 (40)	10 (40)	4	0 (0)
Música instrumental (30)	21 (30)	1 (30)	0	0 (0)
Silencio, siseo y zumbido (30)	11 (30)	4 (30)	0	0 (0)
Voz baja (40)	35,3	24,2	23,5	24,2
Voz de banda telefónica (40)	34,8	25,0	25,6	25,0
Respuestas de una o dos palabras (40)	90,3	22,2	23,6	22,2
Urdu y luego inglés en un clip (20)	—	20,4	44,6	20,4
Voz continua, 25–30 s (15)	—	31,4	32,7	31,4
Inglés, FLEURS (30)	—	5,2	5,8	5,2

Tabla 8: whisper.cpp tal como lo ejecuta una aplicación (q8_0, haz de 5, reintento con temperatura). Filas sin voz: clips con texto inventado (entre paréntesis: con alguna palabra); filas con voz: WER %. «Vacía permitida»: suppress_blank desactivado. La ejecución del modelo base se detuvo tras seis conjuntos al apagarse nuestro servidor de GPU.

Modelo	Indicación	WER (%)	Bucles	Nombres bien
Urdu Sukhan ASR v1	sin indicación	11,7	0	12 de 23
Urdu Sukhan ASR v1	con nombres	11,3	0	17 de 23
Mezcla R4+R5	sin indicación	10,0	0	14 de 23
Mezcla R4+R5	con nombres	9,5	0	17 de 23
Alumno S1 (no publicado)	sin indicación	17,7	0	10 de 23
Alumno S1 (no publicado)	con nombres	32,3	1	9 de 23

Tabla 9: Capítulo de prueba del segundo audiolibro en whisper.cpp, con y sin una indicación que enumera sus nombres.

prefiere para leer. Difiere del estilo de la casa de las transcripciones de los audiolibros, que escriben los préstamos ingleses en escritura urdu; los dos criterios aún no se han unificado en el texto de entrenamiento.

6.8. Dos decodificadores, dos respuestas

El mismo modelo puntúa distinto con los dos decodificadores. La ronda 5 obtuvo un 149,3 % con hablantes no vistos de Common Voice con decodificación voraz en Transformers y un 27,1 % en whisper.cpp, cuyo reintento con temperatura vuelve a decodificar una salida en bucle. Las cifras publicadas deberían nombrar el decodificador; nosotros damos las dos.

7. Velocidad y despliegue

Dónde se va el tiempo. El modelo publicado tiene exactamente la estructura de Whisper large-v3-turbo, así que su velocidad es la del modelo base. En whisper.cpp el codificador domina el tiempo de CPU: siempre procesa una ventana completa de 30 segundos y, para un clip típico, supone alrededor del 95 % del trabajo. Un clip de diez segundos con haz de 5 necesita unos 200 pasos de símbolo del haz, alrededor de 0,8s de decodificación en el Xeon antiguo frente a 22s de codificación. Cada clip paga una ventana completa, así que una respuesta de un segundo cuesta tanto tiempo de codificador como 30 segundos de voz.

Máquinas. La figura 6 da el tiempo del codificador por ventana en cuatro máquinas. La GPU del portátil codifica una ventana en 1,3s; contando la carga del modelo, los conjuntos de prueba de clips de libro (unos 17 segundos cada uno) iban de 4 a 6 veces más rápido que el tiempo real y los de clips de Common Voice de 4 segundos entre 0,8 y 1,7 veces, y redactar los 305 minutos de voz de los audiolibros llevó 59 minutos. Las CPU son mucho más lentas: 5,4s en la CPU del portátil, 11,2s en el servidor en la nube y 21,1s en el Xeon de 2012, que no tiene AVX2 y representa mal a un servidor moderno.

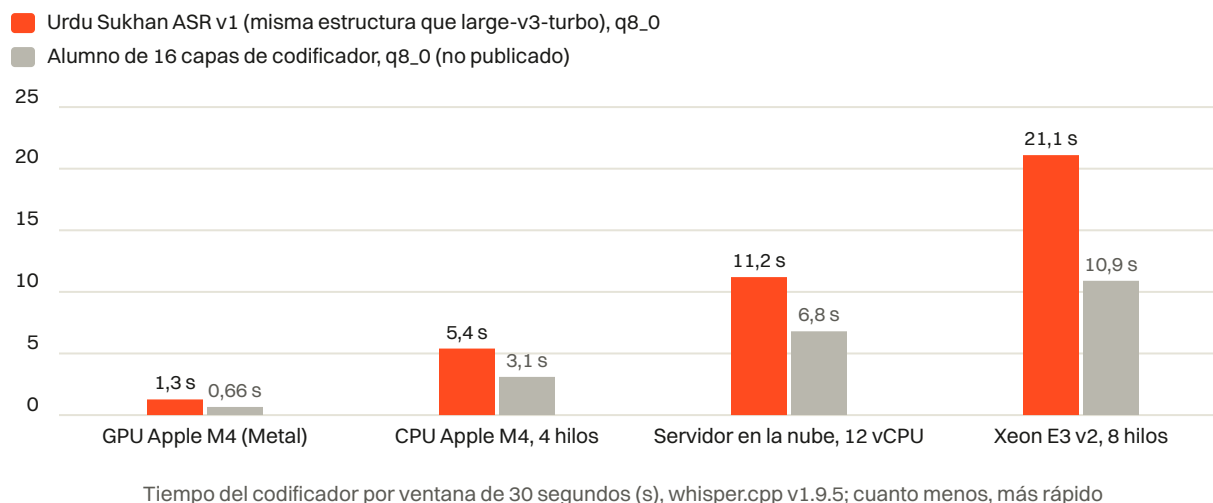


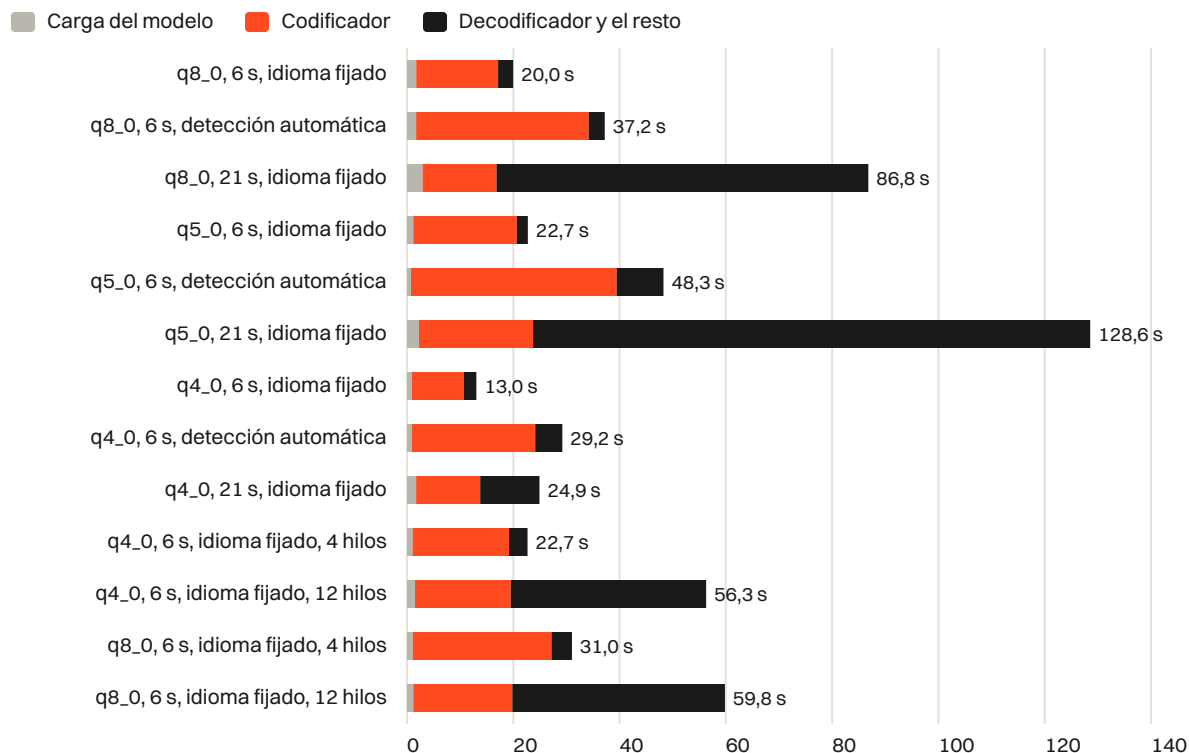
Figura 6: Tiempo del codificador por ventana de 30 segundos, el modelo publicado y el alumno comprimido (sección 8), whisper.cpp v1.9.5.

Un clip en el servidor en la nube. La tabla 10 y la figura 7 miden el modelo publicado con un clip reservado de 6 segundos y otro de 21 segundos en el servidor en la nube, la más rápida de tres ejecuciones, mientras sus otros servicios funcionaban (carga media de 10 a 15 en 12 núcleos). Cuatro resultados importan para las aplicaciones. Primero, *la detección automática del idioma ejecuta el codificador dos veces*: q8_0 tardó 37,2s en lugar de 20,0s con el clip corto. Segundo, *q4_0 es el formato más rápido* en este procesador con AVX2 (codificador 9,7s frente a 15,4s con q8_0) y q5_0 el más lento (19,5s); en el procesador ARM del portátil, q5_0 es algo más rápido que q8_0. Tercero, *los hilos tienen que dejar sitio al resto de la máquina*: con los 12 hilos, el clip corto tardó 59,8s en lugar de 20,0s, casi todo en el decodificador, y con 4 hilos 31,0s; en el mismo servidor con menos carga, 8 y 12 hilos habían sido igual de rápidos. Cuarto, *el reintento con temperatura puede multiplicar la decodificación*: con el clip largo, q8_0 y q5_0 rechazaron su primer resultado y volvieron a decodificar a temperaturas más altas (6600 pasos de símbolo frente a 1096 con q4_0) antes de quedarse con el mismo texto, y tardaron 86,8s y 128,6s frente a 24,9s. Antes, con el servidor más tranquilo, el mismo archivo q8_0 tardó 23,2s con un clip de 5,8 segundos y detección automática (dos pasadas del codificador de 11,2s), así que con los otros servicios ocupados cada pasada del codificador tardó alrededor de un 45 % más.

Formato	Clip (s)	Idioma	Hilos	Carga	Codificación (pasadas)	Decodificación	Total	× tiempo real
q8_0	6	fijado (ur)	8	1,7	15,4 (1)	2,6	20,0	0,30
q8_0	6	automático	8	1,7	32,5 (2)	2,8	37,2	0,16
q8_0	21	fijado (ur)	8	2,9	13,9 (1)	69,6	86,8	0,24
q5_0	6	fijado (ur)	8	1,2	19,5 (1)	1,9	22,7	0,26
q5_0	6	automático	8	0,7	38,7 (2)	8,6	48,3	0,12
q5_0	21	fijado (ur)	8	2,2	21,5 (1)	104,4	128,6	0,16
q4_0	6	fijado (ur)	8	0,9	9,7 (1)	2,1	13,0	0,46
q4_0	6	automático	8	0,9	23,2 (2)	4,9	29,2	0,21
q4_0	21	fijado (ur)	8	1,7	12,1 (1)	10,9	24,9	0,85
q4_0	6	fijado (ur)	4	1,1	18,1 (1)	3,2	22,7	0,27
q4_0	6	fijado (ur)	12	1,5	18,0 (1)	36,4	56,3	0,11
q8_0	6	fijado (ur)	4	1,1	26,1 (1)	3,6	31,0	0,19
q8_0	6	fijado (ur)	12	1,2	18,6 (1)	39,4	59,8	0,10

Tabla 10: El modelo publicado en whisper.cpp v1.9.5 en el servidor en la nube (12 núcleos AMD EPYC compartidos; haz de 5, reintento con temperatura, sin marcas de tiempo); segundos, la más rápida de tres ejecuciones. «Decodificación» incluye el muestreo; entre paréntesis, las pasadas del codificador.

Grabaciones largas. En el servidor en la nube transcribimos una grabación de 27,4 minutos formada por 118 clips reservados con faster-whisper (pesos int8 de CTranslate2, trozos detectados por el detector de voz en lotes de ocho, ocho hilos). Tardó 15,6 minutos con decodificación voraz (1,76 veces el tiempo real, 14,2 % de WER) y 18,1 minutos con haz de 5 (1,52 veces, 13,3 %). Esperábamos que CTranslate2 fuera de dos a cuatro veces más rápido que whisper.cpp, como suele serlo en procesadores Intel, para cuyas instrucciones de 8 bits están afinados sus núcleos de cálculo; los núcleos AMD compartidos a 2,0 GHz de este servidor son el límite. No ejecutamos whisper.cpp con este archivo. A partir del tiempo del codificador por ventana en el servidor tranquilo estimamos de 12 a 13 minutos con q8_0 y de 9 a 10 con q4_0, más o menos lo mismo que faster-whisper; con la carga de la tabla 10, whisper.cpp q4_0 solo alcanzó 0,85 veces el tiempo real con un único clip de 21 segundos. En cualquier caso, una grabación de dos horas tarda allí una hora o más.



Segundos por clip, whisper.cpp v1.9.5, búsqueda en haz de 5, 8 hilos salvo indicación; cuanto menos, más rápido
 Clips reservados de 6 s y 21 s en un servidor en la nube con 12 núcleos AMD EPYC compartidos; otros servicios funcionaban a la vez (carga media 10–15). La más rápida de tres ejecuciones.

Figura 7: Dónde se va el tiempo con un clip en el servidor en la nube.

Tamaños y formatos. Los archivos de whisper.cpp ocupan 874 MB (q8_0), 574 MB (q5_0) y 474 MB (q4_0), y el modelo int8 de CTranslate2 824 MB. Un codificador de Core ML para el Neural Engine de Apple se convirtió sin problemas para el alumno de 16 capas (0,22 s por ventana), pero el codificador completo de 32 capas solo coincidía con PyTorch con una similitud del coseno de 0,984 en fp16, así que no se publica. No medimos el uso de memoria y el modelo aún no se ha ejecutado en un iPhone.

Ajustes recomendados. Fija el idioma cuando lo sepas, o detéctalo una vez por conversación o archivo; desactiva la supresión de la respuesta vacía; pasa antes un detector de voz salvo cuando importe la voz débil; decodifica sin marcas de tiempo y sin arrastrar contexto; en servidores x86 usa q4_0 y unos ocho hilos por trabajo, con varios trabajos en paralelo.

8. Un experimento de compresión

Para que el modelo fuera más rápido en CPU probamos un alumno con menos capas, destilado del modelo publicado. La velocidad solo depende de la estructura, así que primero medimos subconjuntos de capas sin entrenar (tabla 11): el tiempo del codificador es lineal en el número de capas, mientras que las capas del decodificador cuestan poco (2 en lugar de 4 ahorran alrededor de 1,4 ms por paso de símbolo del haz frente a 11 s de codificación). Conservamos las 4 capas del

decodificador y, con ellas, las cabezas de tiempos por palabra de whisper.cpp. whisper-small no era una opción: su codificador solo es alrededor de un 30 % más rápido que el de un alumno turbo de 8 capas, su decodificador de 12 capas es más lento y su urdu es mucho más flojo.

Estructura	Capas enc / dec	Codificador (s)		Decodificador ms/símbolo	GPU M4 cod. (s)	Tamaño (MB)	
		8 hilos	4 hilos			q8_0	q5_0
Modelo completo (32/4)	32 / 4	22,1	41,1	3,94	1,3–1,8	874	574
16/4	16 / 4	11,1	20,2	3,89	0,85	539	357
16/2	16 / 2	10,7	21,9	2,5	–	483	321
12/2	12 / 2	8,2	16,0	2,41	0,58	399	266
8/4	8 / 4	5,7	11,6	3,79	–	371	248
8/2	8 / 2	5,5	10,8	2,44	0,38	315	212
whisper-small	12 / 12	4,0	7,9	4,77	–	264	175
whisper-base	6 / 6	1,2	2,3	1,63	–	82	55

Tabla 11: Velocidad y tamaño de subconjuntos de capas sin entrenar del modelo publicado y de dos modelos Whisper de serie (whisper-bench, Xeon E3 v2; una ventana de 30 segundos; coste del decodificador por paso de símbolo del haz).

Dos pruebas piloto de 600 pasos compararon 16 y 8 capas de codificador: el alumno de 8 capas seguía por encima del 100 % de WER en todas las fuentes reservadas, así que entrenamos el de 16 (una capa de cada dos, conservando la primera y la última). La pérdida combinaba la entropía cruzada con las transcripciones, una divergencia KL respecto de la distribución del maestro para el siguiente símbolo a temperatura 2 y el error cuadrático medio respecto de la salida del codificador del maestro y de las salidas normalizadas de las capas del maestro de las que se copió cada capa del alumno, con SpecAugment solo en el alumno. Los clips de libro y sin voz se entrenaron solo con sus transcripciones, para que el alumno no copiara los nombres mal escritos del maestro. Cada clip se llevó a un nivel normal antes del aumento de datos (sección 6.5) y se empaquetaron de dos a cuatro clips cortos en cada ventana de 30 segundos.

Dos problemas obligaron a reiniciar (figura 8). El alumno escribía una palabra en cada clip de silencio, así que multiplicamos por 40 el peso del símbolo de fin de texto en las ventanas sin voz; sus respuestas en clips sin voz empezaron entonces a variar entre puntos de control, así que añadimos «se forzó una palabra primero, ahora para»: en la mitad de las ventanas sin voz se coloca una palabra común en la entrada del decodificador sin puntuarla, y solo se puntúa el fin de texto que la sigue. Cada reinicio rompió brevemente el comportamiento sin voz, que se recuperó en el siguiente punto de control. Tras 7200 pasos (unas 11 horas de GPU), el alumno conservado en el paso 6600, S1, era 1,9 veces más rápido en el codificador en todas las máquinas e inventaba texto en 0, 1 y 0 clips de ruido, música y silencio en whisper.cpp, frente a 10, 1 y 4 del modelo publicado. Pero quedó entre 0,5 y 5,4 puntos por detrás en los conjuntos de prueba y era flojo en inglés (tabla 12), y los últimos 3000 pasos ganaron alrededor de un punto. Quitar la mitad de un codificador de 32 capas descarta más de lo que 110 horas de audio pueden recuperar: el camino es más audio transcrito por el maestro, no más pasos. El autor probó los dos modelos en el servidor en la nube y prefirió el modelo publicado; S1 no se publica.

WER (%)	Modelo completo	Alumno S1	Diferencia
IndicVoices urdu, validación	12,7	17,2	+4,5
Common Voice, hablantes no vistos	28,1	30,9	+2,8
FLEURS urdu, prueba	16,2	21,3	+5,1
UrduSpeech, partición de prueba	8,8	10,8	+2,0
nzg73, clips reservados	1,6	2,1	+0,5
Audiolibro 1, capítulos reservados	4,0	6,6	+2,6
Audiolibro 2, capítulo reservado	13,2	18,6	+5,4
Voz baja (whisper.cpp)	24,2	28,8	+4,6
Voz de banda telefónica (whisper.cpp)	25,0	25,9	+0,9
Respuestas de una o dos palabras (whisper.cpp)	22,2	23,6	+1,4
Urdu y luego inglés en un clip (whisper.cpp)	20,4	38,8	+18,4
Voz continua, 25–30 s (whisper.cpp)	31,4	34,8	+3,4
Inglés, FLEURS (whisper.cpp)	5,2	23,0	+17,8

Tabla 12: El modelo publicado y el alumno de 16 capas S1. Conjuntos de prueba: Transformers, voraz; pruebas de estrés: whisper.cpp q8_0 sin la supresión de la respuesta vacía.

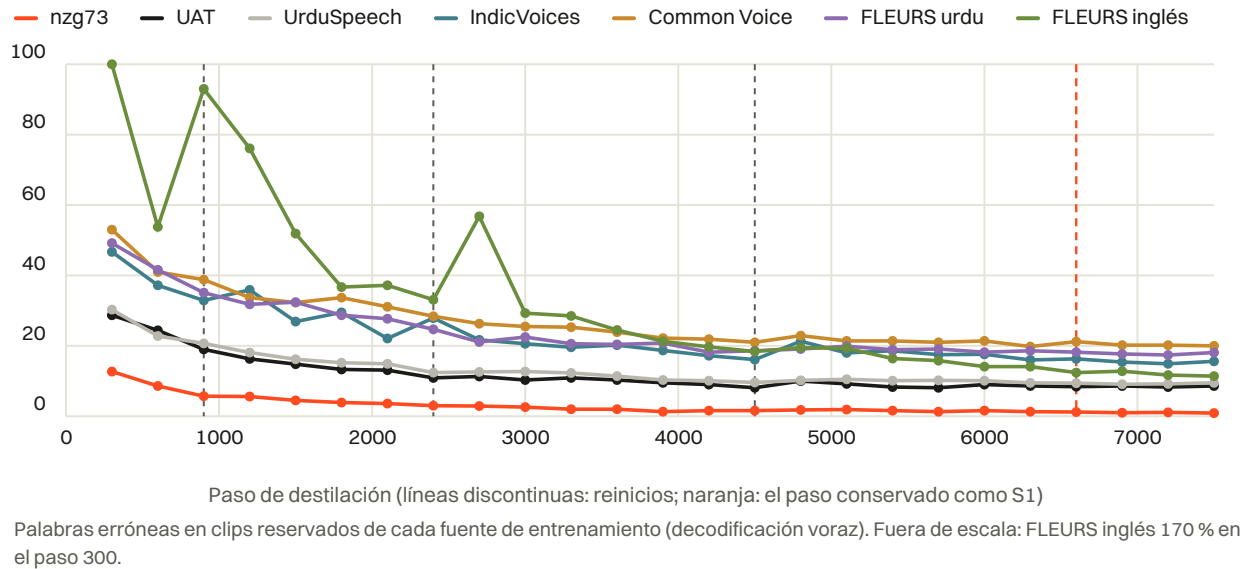


Figura 8: Palabras erróneas del alumno en clips reservados de cada fuente de entrenamiento durante la destilación.

9. Lecciones

Cuatro lecciones de este trabajo valen más allá del urdu. Primera, *el código de evaluación necesita tanto cuidado como el de entrenamiento*: un archivo de prueba con filas duplicadas, una regla de bucles que cambió tres veces, una regla de dos palabras que ocultaba las respuestas de una sola y dos trabajos de puntuación que borraban los resultados del otro y mostraron un falso fallo cambiaron cifras que ya habíamos dado. Segunda, *el nivel del audio forma parte de la etiqueta*: voz débil con transcripción completa enseñó a dos rondas a transcribir el siseo. Tercera, *los valores por defecto del decodificador forman parte del modelo*: las marcas de tiempo, la supresión de la respuesta vacía, la detección del idioma y el reintento con temperatura cambiaron los resultados más que algunas decisiones de entrenamiento. Cuarta, *mide la velocidad en la máquina de destino*: el procesador de 2012 del servidor de GPU era varias veces más lento que el servidor en la nube y daba una imagen equivocada, y en el servidor en la nube la carga de otros servicios cambió qué número de hilos era el más rápido.

10. Limitaciones y amenazas a la validez

Los conjuntos de prueba de los audiolibros son pequeños (58 clips, el segundo de solo 2,9 minutos) y proceden de los mismos dos narradores cuyos otros capítulos se usaron para entrenar; los de respuestas cortas y de mezcla tienen 40 y 20 clips. No damos intervalos de confianza. IndicVoices es sobre todo urdu de la India, las mujeres están infrarrepresentadas en la prueba de Common Voice y no tenemos ninguna prueba de conversación espontánea grabada por teléfono por hablantes de Pakistán. La ronda 6 oyó los archivos de MUSAN de las pruebas de estrés de ruido y música durante el aumento de datos, así que sus resultados ahí son algo favorables. Las causas que damos para los fallos de la ronda 5 y del último paso de la ronda 6 encajan con las mediciones, pero no se probaron una a una. Las rondas 1 a 4 se midieron con conjuntos de prueba que cambiaron entre rondas, y algunas cifras tempranas de Common Voice proceden de un archivo con duplicados. Nunca puntuamos el propio Whisper large-v3, ni medimos la memoria, ni ejecutamos el modelo en un teléfono. Los tiempos del servidor en la nube varían con los otros servicios que funcionan en él. La ejecución de estrés del modelo base en whisper.cpp está incompleta.

11. Reproducibilidad y disponibilidad

El código (herramienta de revisión, estilo de la casa, preparación de datos, entrenamiento, destilación, evaluación, exportaciones y la página de pruebas), la tabla de resultados de cada evaluación y los scripts que generan las tablas y figuras de este artículo a partir de ellos están en el repositorio de Urdu Sukhan ASR, <https://github.com/Loudness-ai/urdu-sukhan-asr>. El modelo está publicado en Hugging Face como Loudness-ai/urdu-sukhan-asr-v1 (<https://huggingface.co/Loudness-ai/urdu-sukhan-asr-v1>) en formatos Transformers, whisper.cpp y CTranslate2 con la licencia CC BY-NC 4.0 para uso no comercial; el acceso se concede bajo solicitud. El punto de control original de precisión completa se perdió al apagar nuestro servidor de entrenamiento al final del trabajo; los pesos de Transformers se reconstruyeron a partir del archivo q8_0 y se comprobaron con los resultados de este artículo. La instalación y los ajustes están en <https://loudness.ai/es/developers/urdu-sukhan-asr/>. Las grabaciones de los audiolibros y el audio de UAT transcrito no se redistribuyen. Nuestros conjuntos de prueba de Common Voice, FLEURS y de estrés están disponibles bajo solicitud en <https://huggingface.co/datasets/Loudness-ai/urdu-sukhan-asr-test-sets>. Los conjuntos

públicos están disponibles en sus editores con sus propias condiciones. Este informe está archivado en Zenodo, <https://doi.org/10.5281/zenodo.23518703>, con la licencia CC BY 4.0.

12. Conflictos de intereses

El autor desarrolla Loudness.ai, cuyos productos usarán el modelo descrito aquí. No se recibió financiación externa.

Historial de versiones. Versión 1.0 (21 de octubre de 2026): primera publicación.

Referencias

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey e I. Sutskever. Robust speech recognition via large-scale weak supervision. En *Proc. ICML*, 2023.
- [2] S. Arif, A. J. Khan, M. Abbas, A. A. Raza y A. Athar. WER we stand: benchmarking Urdu ASR models. arXiv:2409.11252, 2024.
- [3] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann y M. Sloane. Careless Whisper: speech-to-text hallucination harms. En *Proc. ACM FAccT*, 2024.
- [4] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera y A. Bapna. FLEURS: few-shot learning evaluation of universal representations of speech. En *Proc. IEEE SLT*, 2023.
- [5] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers y G. Weber. Common Voice: a massively-multilingual speech corpus. En *Proc. LREC*, 2020.
- [6] T. Javed et al. IndicVoices: towards building an inclusive multilingual speech dataset for Indian languages. En *Findings of ACL*, 2024.
- [7] D. Snyder, G. Chen y D. Povey. MUSAN: a music, speech, and noise corpus. arXiv:1510.08484, 2015.
- [8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk y Q. V. Le. SpecAugment: a simple data augmentation method for automatic speech recognition. En *Proc. Interspeech*, 2019.
- [9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang y W. Chen. LoRA: low-rank adaptation of large language models. En *Proc. ICLR*, 2022.
- [10] M. Wortsman et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. En *Proc. ICML*, 2022.
- [11] G. Hinton, O. Vinyals y J. Dean. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
- [12] S. Gandhi, P. von Platen y A. M. Rush. Distil-Whisper: robust knowledge distillation via large-scale pseudo labelling. arXiv:2311.00430, 2023.
- [13] G. Gerganov y colaboradores. whisper.cpp, versión 1.9.5. <https://github.com/ggml-org/whisper.cpp>.
- [14] SYSTRAN. faster-whisper, versión 1.2.1. <https://github.com/SYSTRAN/faster-whisper>.
- [15] Silero Team. Silero VAD: pre-trained enterprise-grade voice activity detector. <https://github.com/silero4/silero-vad>.
- [16] G. Kunz y colaboradores. Chartist, versión 2.1.0. <https://chartist.dev>.