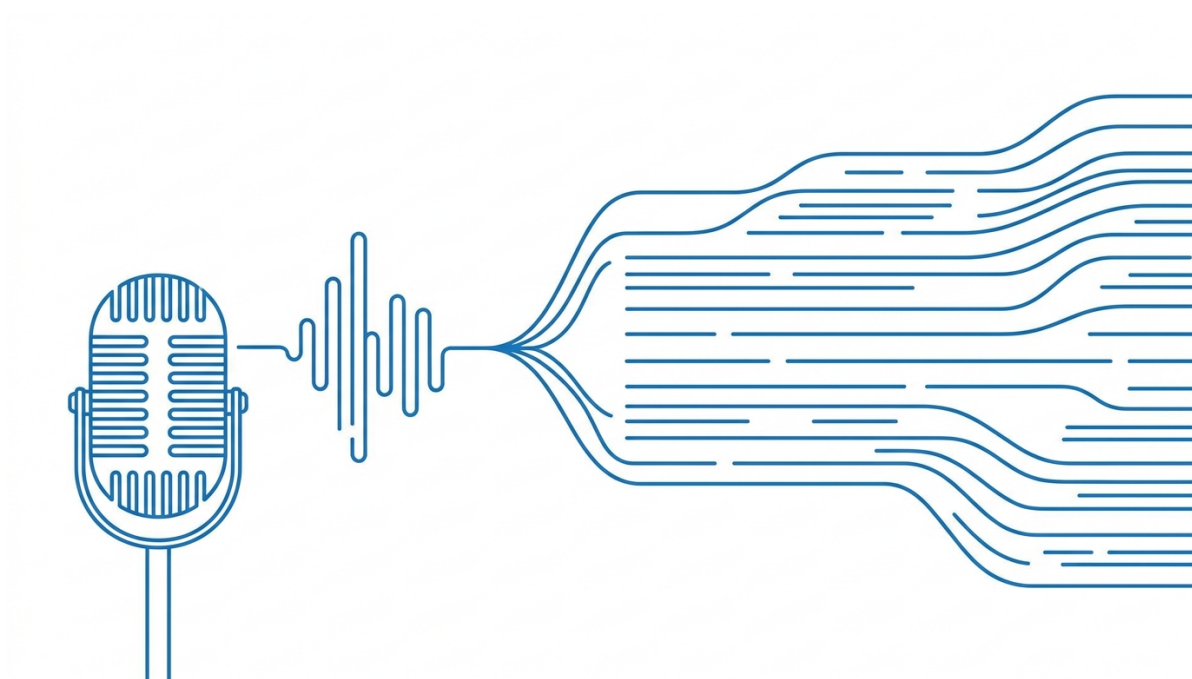


LO ESENCIAL · Módulo 8 · Voz y multimodal

Transcribir, separar hablantes y sintetizar en local. Y auditar qué herramientas «locales» mandan datos afuera.

8.1 · Transcripción y diarización local



Voz y multimodal

Objetivos

- Producir un acta con hablantes de una reunión
- Elegir tamaño de modelo por latencia y calidad

De la reunión al acta

La transcripción local convierte audio en texto; la **diarización** dice quién habló cuándo. Con las dos, una reunión de comité de 40 minutos se vuelve un borrador de acta en minutos, sin que el audio salga de la oficina.

Piezas [VOLÁTIL – verificado: 2026-09]

Pieza	Criterio de elección	Pe
Modelo de transcripción (familia Whisper y derivados optimizados)	español, tamaño vs. calidad, velocidad en CPU	A:
Diarización (pyannote o similar)	número de hablantes esperado; licencia	B/
Alineación palabra a palabra	necesaria si el acta cita tiempos	op

Criterio de tamaño: el modelo más pequeño cuyo WER (tasa de error por palabra) en **una grabación real del cliente** sea aceptable. Se mide igual que todo: 10 minutos transcritos a mano contra el modelo.

PROCEDIMIENTO · Laboratorio 8.1 · acta de una reunión de 10 minutos

```
cd ~/labs/el-tornillo && mkdir -p labs/08-01 && cd labs/08-01 && pip install -q faster-
whisper
```

```
# transcribir.py – transcripción local con marcas de tiempo
from faster_whisper import WhisperModel
m = WhisperModel("<tamaño>", device="auto", compute_type="int8") # [VOLÁTIL] small |
medium | large-v3
segs, info = m.transcribe("reunion.wav", language="es", vad_filter=True)
with open("transcripcion.txt", "w") as f:
    for s in segs: f.write(f"[{s.start:7.1f}-{s.end:7.1f}] {s.text.strip()}\n")
print("idioma:", info.language, "· duración:", round(info.duration), "s")
```

Con diarización, cada línea lleva además **HABLANTE_01**. Después, el asistente (5.2) produce el acta con esquema: asistentes, orden del día, decisiones, responsables, fechas. Lo que debe ver: un acta que el gerente **corrige**, no que redacta.

Entregable: transcripcion.txt con hablantes, el acta generada, y la medición de WER sobre 10 minutos. Commit: «Capa 8.1: transcripción».

Qué se degrada en cada perfil

- A** — Modelo pequeño; 40 min de audio tardan ~20–40 min en CPU; diarización lenta
- B** — Modelo medio; tiempo real o mejor
- C** — Modelo grande; varias reuniones en paralelo

Cuándo NO usar esto Transcripción y diarización local

- No transcriba reuniones sin avisar a los asistentes: la política de datos (2.3) y la ley lo exigen.

- No confíe en los nombres: la diarización dice «hablante 1», no «Carlos». El mapeo lo hace una persona.

PRÁCTICA · Práctica

1. Calcule el WER de su transcripción con 10 minutos anotados a mano.
2. ¿Qué campos del esquema del acta necesitan confianza y cuáles no?

Referencias

1. Radford, A. et al. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv:2212.04356 (Whisper).
2. faster-whisper — <https://github.com/SYSTRAN/faster-whisper> · pyannote.audio — <https://github.com/pyannote/pyannote-audio> [VOLÁTIL – verificado: 2026-09]

8.2 · Síntesis de voz local

Objetivos

- Responder por voz desde el asistente
- Comparar voces por naturalidad y costo

Responder por voz

En el mostrador de la ferretería, el vendedor no puede leer una pantalla mientras atiende. Una respuesta corta **por voz** («quedan 40 unidades, el proveedor entrega el jueves») sí sirve. La síntesis local convierte texto en audio sin salir de la red.

Criterios [VOLÁTIL – verificado: 2026-09]

Criterio	Por qué
Español con acento aceptable	las voces «neutras» suenan extranjeras; se prueba con el cliente
Latencia < 1 s para frases cortas en CPU	si tarda más, nadie la usa
Licencia libre para uso comercial	va en un producto
Tamaño ≤ 100 MB por voz	corre en el nodo de servicios

Motores ligeros (familia Piper y similares) cumplen los cuatro en el perfil A. Motores más naturales exigen GPU y se justifican solo si la voz es parte del producto.

PROCEDIMIENTO · Laboratorio 8.2

```
cd ~/labs/el-tornillo && mkdir -p labs/08-02 && cd labs/08-02
echo "Quedan cuarenta unidades. El proveedor entrega el jueves." | piper --model <voz-es>
--output_file r.wav
afplay r.wav 2>/dev/null || aplay r.wav
```

Conecte la salida del asistente (respuesta corta, ≤ 30 palabras, sin cifras largas: «mil ochocientos cincuenta mil» se dicta mal) a la síntesis. Mida latencia texto→audio.

Entregable: responder_voz.py, tres audios de ejemplo, latencias. Commit: «Capa 8.2: VOZ».

Qué se degrada en cada perfil

A — Motores ligeros en tiempo real; voces más naturales, no

B — Motores neurales más naturales en vivo

Cuándo NO usar esto Síntesis de voz local

- No dicte cifras de dinero completas ni números de documento: se muestran en pantalla.
- No use síntesis en la nube «porque suena mejor»: el texto de la respuesta contiene datos del cliente. Lección siguiente.

PRÁCTICA · Práctica

1. Reescriba «El total es \$1 850 000 y vence el 2026-10-03» para que se dicte bien.
2. ¿Qué respuestas del asistente **no** deberían tener versión de voz?

Referencias

1. Piper — <https://github.com/rhasspy/piper> [VOLÁTIL – verificado: 2026-09]
2. Módulo 2.3 — datos personales en audio.

8.3 · Lo que parece local y no lo es

Objetivos

- Auditar el tráfico de red de una herramienta
- Escribir el informe de egreso de datos

La advertencia

Muchas herramientas se presentan como «locales» o «privadas» y **mandan datos afuera**: telemetría con fragmentos de texto, «mejoras» que suben documentos, modelos que en realidad llaman a una API, extensiones de escritorio que sincronizan «para conveniencia». No es paranoia: es lo que hay que verificar antes de ponerle un contrato de confidencialidad encima.

La única prueba válida es **mirar el tráfico**.

PROCEDIMIENTO · Laboratorio 8.3 · auditar el egreso de cinco herramientas

```
cd ~/labs/el-tornillo && mkdir -p labs/08-03 && cd labs/08-03
# 1) Captura del tráfico de un contenedor/proceso durante un uso normal de 10 minutos
sudo tcpdump -i any -w captura.pcap 'not (net 100.64.0.0/10 or net 10.0.0.0/8 or net 192.168.0.0/16)'
# 2) ¿A dónde fue?
tshark -r captura.pcap -T fields -e ip.dst -e tls.handshake.extensions_server_name | sort
| uniq -c | sort -rn | head -20
```

Para cada herramienta (runtime, interfaz de chat, OCR, transcripción, síntesis): destino, cuándo (al arrancar, al usar, periódico), qué (si es TLS, al menos el nombre del servidor), y si se puede **desactivar** por configuración. Verifique también con la red **desconectada**: si deja de funcionar, no era local.

Herramienta	Destinos fuera de la malla	Cuándo	Desactivable	Veredicto
...				local / local con telemetría / no local

La regla de la topología: el nodo de cómputo y el de servicios **no tienen salida a internet** salvo por una lista blanca (repositorios de paquetes, descargas de modelos) que se cierra después de instalar. Así, lo que intente salir, falla —y se ve en el registro.

Entregable: la tabla con 5 herramientas y labs/08-03/lista-blanca.md. Commit: «Capa 8.3: egreso auditado».

Nada. Es igual en A, B y C, y es de lo que más vale ante un cliente.

Cuándo NO usar esto Lo que parece local y no lo es

- No dé por buena la política de privacidad escrita: se audita el tráfico, no el PDF.
- No bloquee todo egreso sin lista blanca: las actualizaciones de seguridad también salen por ahí.

PRÁCTICA · Práctica

1. Una herramienta manda «solo métricas anónimas» a un dominio de analítica. ¿La deja? ¿Con qué condición contractual?
2. Escriba la regla de firewall del nodo de cómputo: sin egreso salvo lista blanca.

Referencias

1. Wireshark / tshark — <https://www.wireshark.org/docs/>
2. OWASP LLM Top 10 — LLM06 (divulgación de información) y LLM03 (cadena de suministro).
3. Ley 1581 de 2012 — transferencia y transmisión de datos a terceros (arts. 26–27).