

IA local para PyMEs · Módulo 8

# Voz y multimodal

---

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

# Qué se construye en este módulo

## **Transcripción y diarización local**

Producir un acta con hablantes de una reunión

## **Síntesis de voz local**

Responder por voz desde el asistente

## **Lo que parece local y no lo es**

Auditar el tráfico de red de una herramienta

---

## 8.1 · Transcripción y diarización local

# Transcripción y diarización local

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.

# En una tabla

| Pieza                                                             | Criterio de elección                 |
|-------------------------------------------------------------------|--------------------------------------|
| Modelo de transcripción (familia Whisper y derivados optimizados) | español, tamaño vs. velocidad en CPU |
| Diarización (pyannote o similar)                                  | número de hablantes, licencia        |
| Alineación palabra a palabra                                      | necesaria si el acta o               |

# El comando, copiable

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

Si no corre desde cero con un comando, no entra al curso.

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.

# Qué se degrada en cada perfil

## A · sin GPU

Modelo pequeño; 40 min de audio tardan ~20–40 min en CPU; diarización lenta

## B · 8–16 GB VRAM

Modelo medio; tiempo real o mejor

## C · 24 GB+

Modelo grande; varias reuniones en paralelo

---

## 8.2 · Síntesis de voz local

# Síntesis de voz local

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.

# En una tabla

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

# El comando, copiable

```
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
```

Si no corre desde cero con un comando, no entra al curso.

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.

---

## 8.3 · Lo que parece local y no lo es

# Lo que parece local y no lo es

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.

# El comando, copiable

```
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
```

Si no corre desde cero con un comando, no entra al curso.

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.

# Entregables del módulo

- `transcripcion.txt` con hablantes, el acta generada, y la medición de WER sobre 10 minutos.
- `responder_voz.py`, tres audios de ejemplo, latencias.
- la tabla con 5 herramientas y `labs/08-03/lista-blanca.md`.

**Transcribir, separar  
hablantes y  
sintetizar en local.**