VOZ A TEXTO Y
TEXTO A VOZ

ESTRUCTURA DE LA CLASE

INTRODUCCIÓN

Presentación del tema de la clase

Repaso sobre procesamiento de audio

RECONOCIMIENTO DE VOZ

Métodos de reconocimiento de sonido

Conversión de texto a voz

Funcionamiento de Whisper

Uso de Whisper

SÍNTESIS DE VOZ

Métodos de síntesis de sonido y voz

Tipos de modelos

Clonación de voz

Presentación de Coqui-TTS

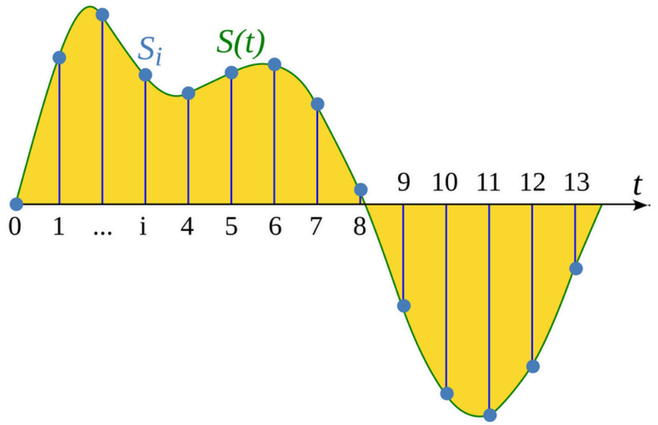
ACTIVIDAD PRÁCTICA

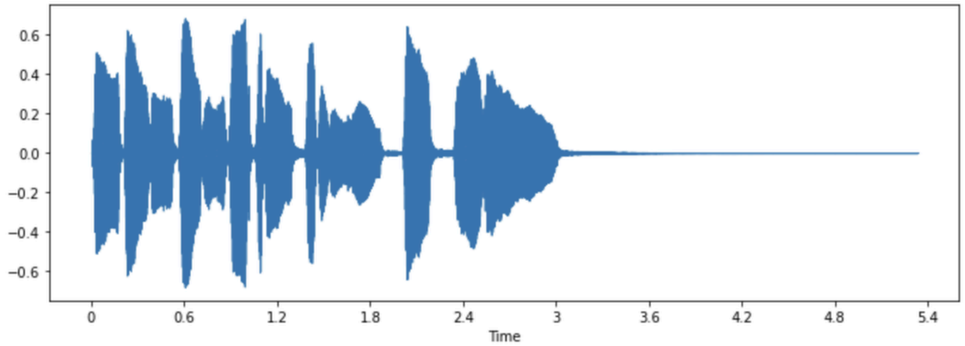
Chatbot voz a voz

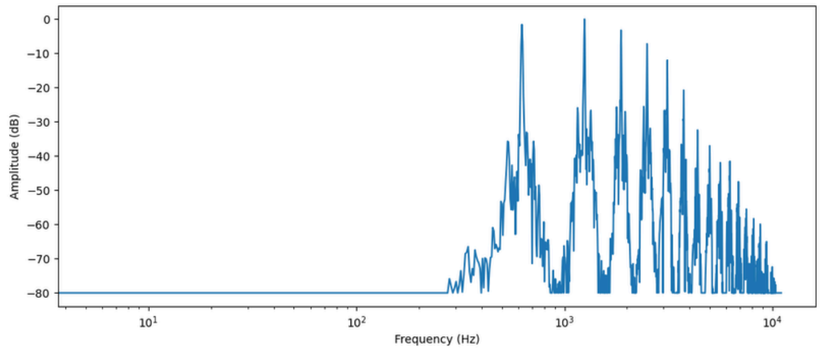
CONCLUSIONES

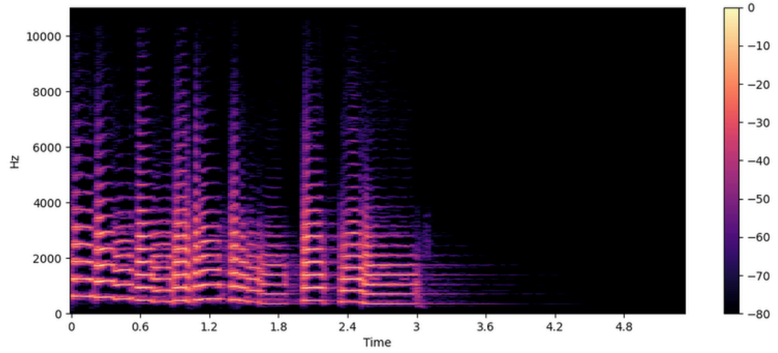
Recapitulación de los puntos clave de la clase

PROCESAMIENTO DE AUDIO

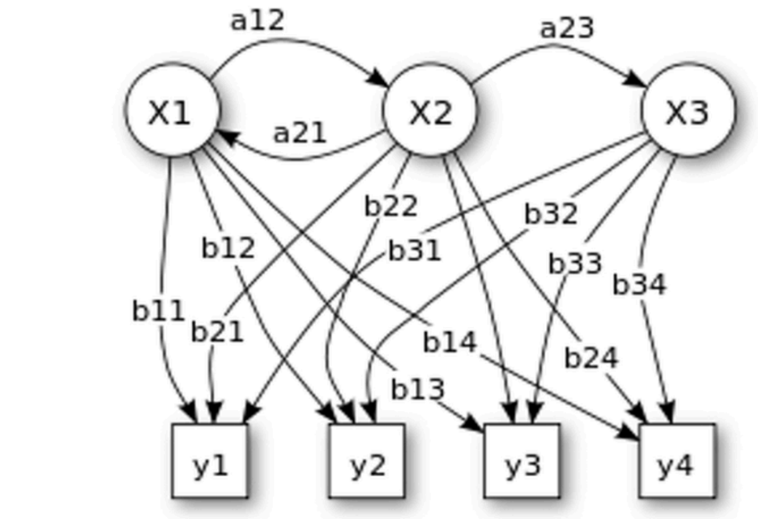








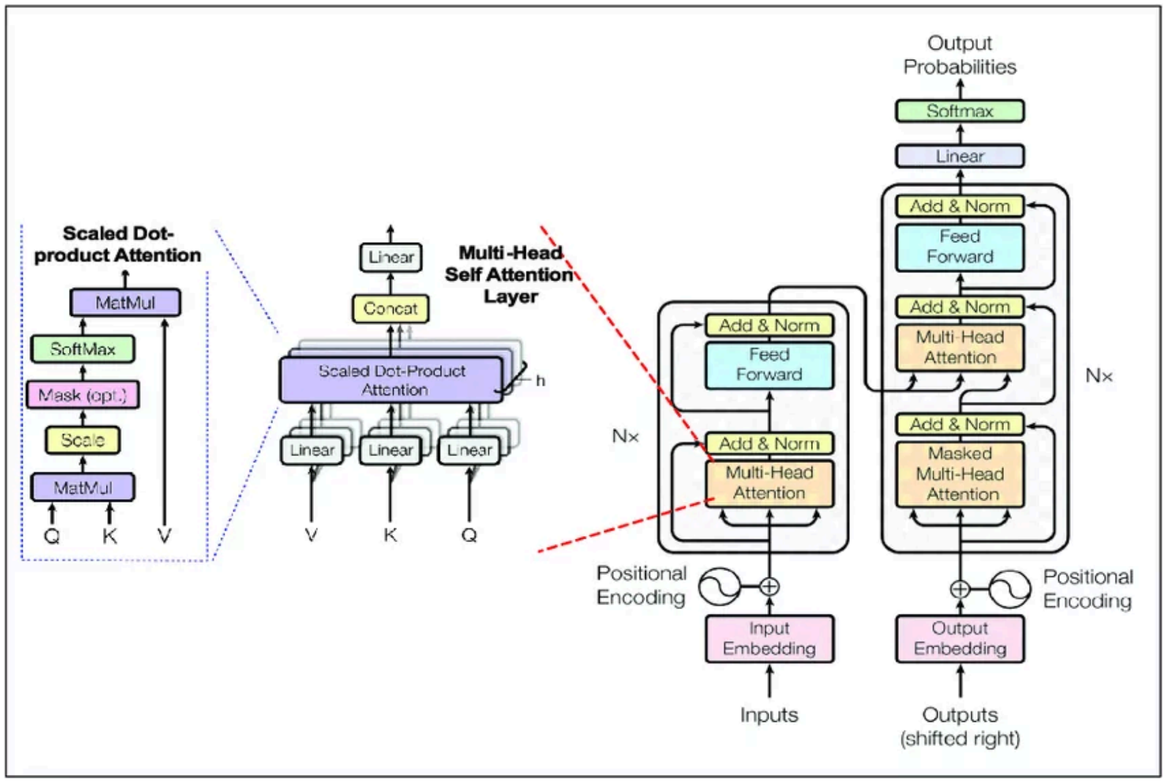
MODELOS ACÚSTICOS



Son el modelo más clásico de SST y su representante principal son los modelos de estado oculto

Las HMM o las NNs entrenan una red con voz y transcripciones para aprender a predecir la palabra que corresponde a una secuencia de sonidos

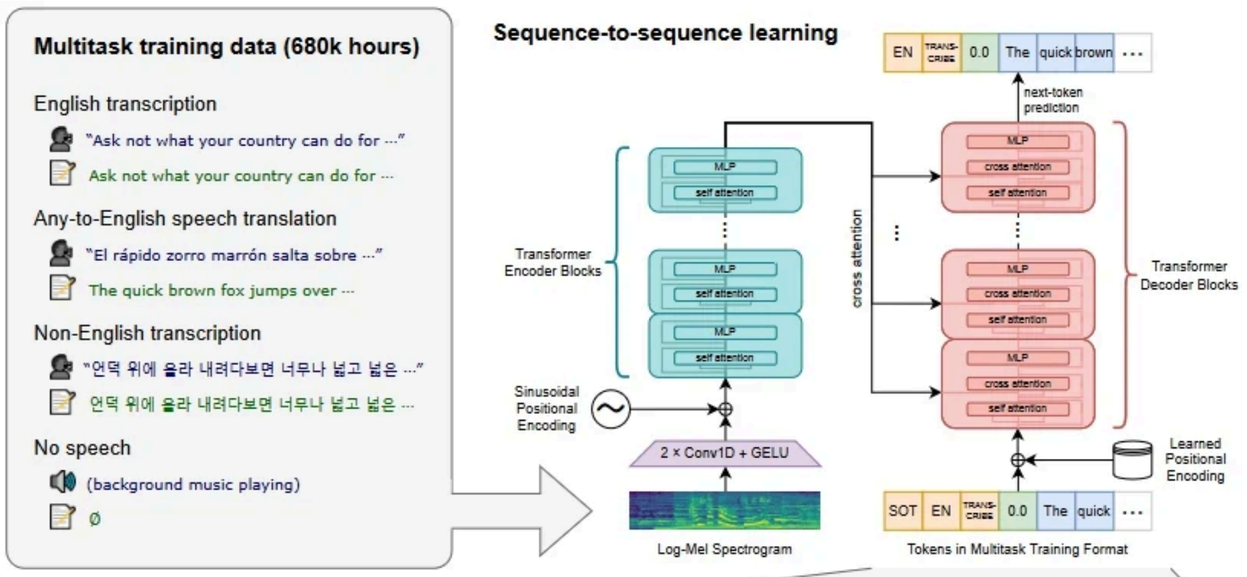
MODELOS DE LENGUAJE

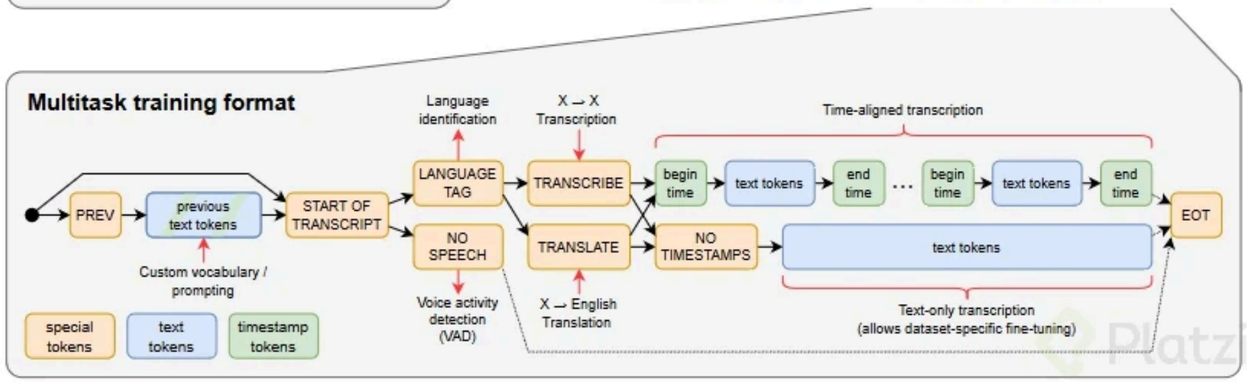


A través de una red neural la red extrae features y dependencias de la voz

Pueden ser basados en n-gramas, redes recurrentes como LSTMs, o basados en contexto como GPTs.

FUNCIONAMIENTO DE WHISPER





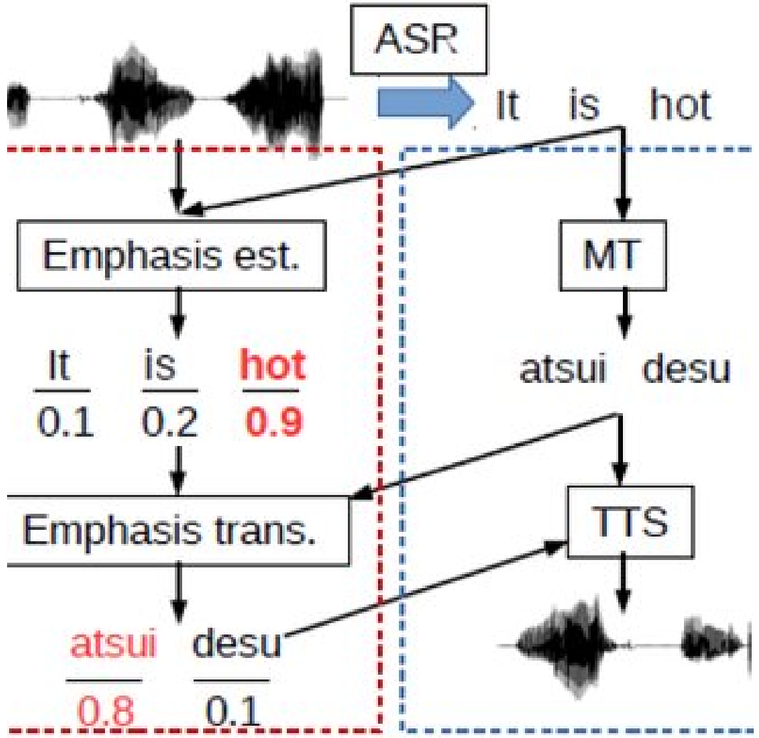
USO DE WHISPER

```python
import whisper

# Transcribimos usando el modelo base de whisper
model = whisper.load_model("base")
result = model.transcribe(<audio>, language=<idioma>)

print(result['text']) # resultado de la transcripción
```

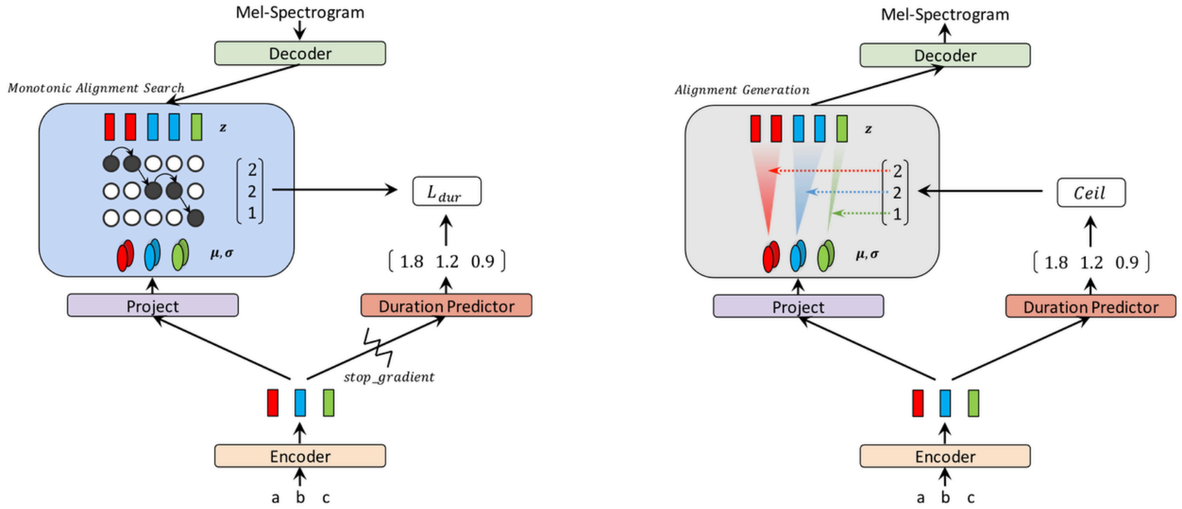
MODELOS CONCATENATIVOS



Son los más antiguos y sintetizan la voz a través de la asignación previa de sonidos a determinados tokens

Los más avanzados son capaces de extraer features y parámetros para mejorar las predicciones

MODELOS POR PARÁMETROS



(a) An abstract diagram of training procedure.

(b) An abstract diagram of inference procedure.

Figure 1. Training and inference procedures of Glow-TTS.

El modelo aprender a generar las ondas de sonido de una fuente base

Puede generar ondas sinusoidales a partir del origen (SWS) o a través de features melspectrales (Vocoder).

Permite alcanzar una voz natural e incluso clonar la voz

MODELOS TTS

ESPECTROGRAMA

Aprende las relaciones entre el texto y su espectrograma

WAVEFORM Y E2E

Toman la fuente de audio, pudiendo o no convertirla en un espectrograma, y aprende a generar sonido a partir del texto dado

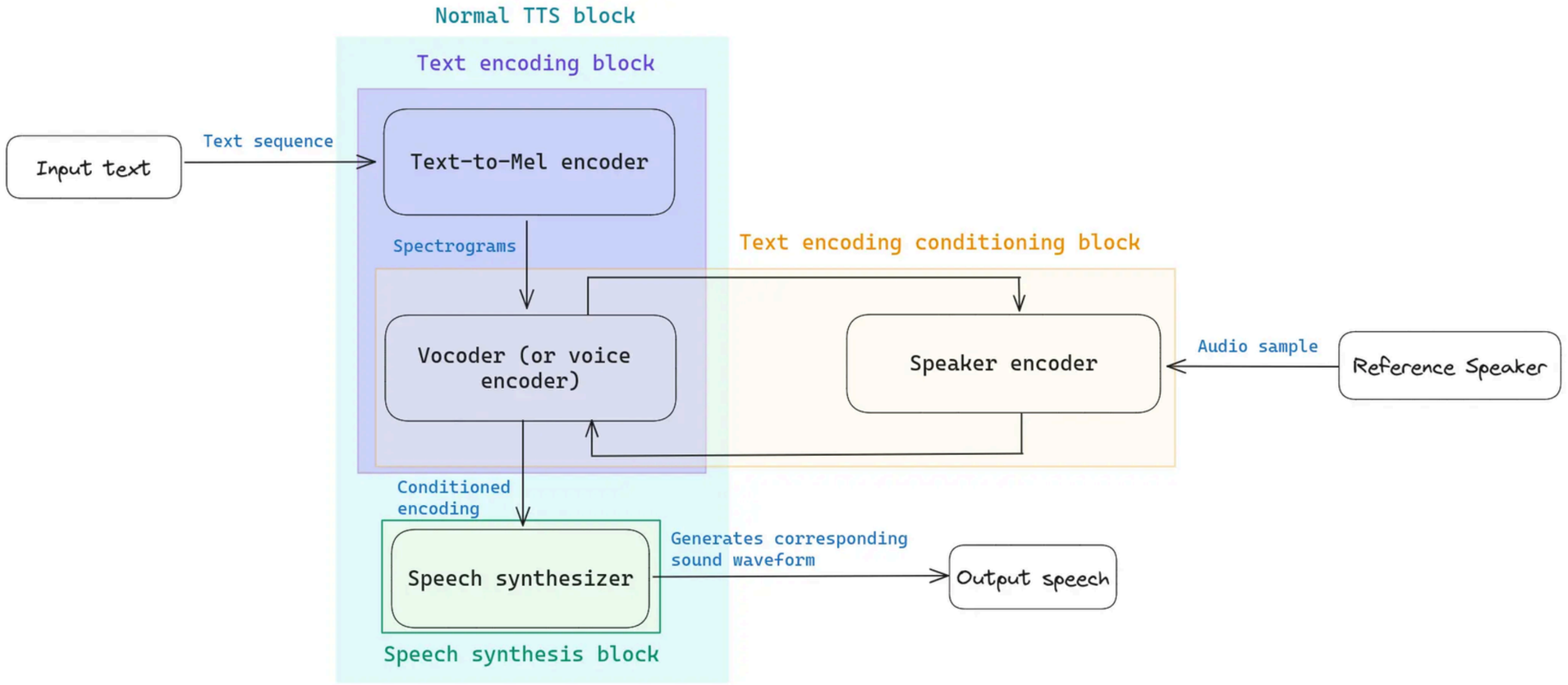
ENCODERS Y VOCODERS

A partir del audio (o espectrogramas) aprenden las características del habla del locutor

VOICE CONVERTION

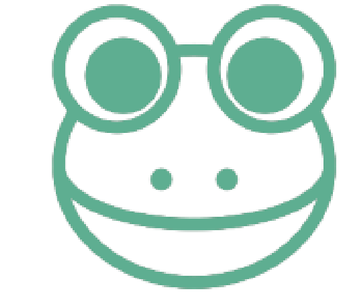
Haciendo uso de uno o más modelos anteriores, este modelo busca cambiar la voz del locutor

CLONACIÓN DE VOZ



Un encoder toma las características de la voz de un locutor de referencia y se las suministra a un codificador de voz cuyos outpus son usados por el sintetizador

COQUI-TTS



TTS

⊗TTS

🌮 Tacotron

🐶 Bark

Coqui dispone de varios modelos TTS que permite la sintetización de voces y la clonación a la vez que entrenar los modelos existentes para lograr mejores resultados

USO DE COQUI-TTS

```
import torch
from TTS.api import TTS

# Get device
device = "cuda" if torch.cuda.is_available() else "cpu"

# Listar todos los modelos disponibles
print(TTS().list_models())

# Inicializar TTS
tts = TTS(<modelo_escogido>).to(device)

# Generar array
wav = tts.tts(text=<texto>, speaker_wav=<voz a clonar>, language=
<idioma>)

# Generar wav
tts.tts_to_file(text=<texto>, speaker_wav=<voz a clonar>, language=
<idioma>)
```

CONCLUSIONES

VOZ A TEXTO

La IA codifica el lenguaje y un decodificador genera una secuencia de texto a partir de esa codificación. Whisper es el modelo más importante a la fecha.

TEXTO A VOZ

El texto se codifica y es pasado a un sintetizador que genera la voz para el conjunto de caracteres dado. CoquiTTS es una de las soluciones OpenSource más populares.