Universidad Intercontinental de la Empresa

Asignatura

Sistemas Interactivos Inteligentes

Práctica 4. Unidad V

TTS

Alumna: Carlota Fernández del Riego

Índice

1. Resumen

La síntesis de voz mediante modelos neuronales es actualmente una de las áreas más importantes dentro de los sistemas interactivos inteligentes. En esta práctica se han evaluado dos modelos de Text-to-Speech (TTS) orientados a la clonación de voz, con el objetivo de generar audios sintéticos a partir de un fragmento de voz real.

El objetivo principal ha sido analizar cómo cada modelo reproduce la identidad vocal, la naturalidad, y la estabilidad de la voz. Para ello se ha realizado una preparación del audio, una generación de archivos sintéticos y una extracción de embeddings para posteriormente calcular similitudes y diferencias.

2. Preparación de Datos

2.1. Recorte del audio original

El audio descargado inicialmente contenía fragmentos innecesarios y varios silencios prolongados por lo que se recortó para conservar únicamente la parte donde la voz es clara y estable.

Este ajuste permitió:

- Reducir ruido y variabilidad en el embedding.
- Mejorar la estimación del **speaker identity vector.**
- Favorecer una prosodia más coherente.
- Evitar que silencios distorsionen el timbre percibido.

El fragmento final, es de aproximadamente 10 segundos, y fue suficiente para que los modelos captasen correctamente las características vocales.

2.2. Procesado y normalización

Se realizó una conversión del audio a formato **WAV**, **16 kHz** y **canal mono**, de acuerdo con los requisitos de ambos modelos, ya que en un primer momento fue descargado en formato mp3. Por otra parte, se aplicó una normalización para asegurar que los niveles de la señal no se saturasen durante el procesado y se verificó la forma de onda con soundfile y librosa para garantizar que no existían problemas de clipping, asegurando así que el audio cumplía con las condiciones necesarias para una evaluación fiable.

3. Modelos Evaluados

3.1. XTTS

XTTS es un modelo TTS multilingüe de arquitectura encoder-decoder que integra tres componentes principales:

1. Un **speaker encoder** entrenado con aprendizaje contrastivo para extraer representaciones robustas del timbre.
2. Un **encoder lingüístico multilingüe** capaz de generalizar entre distintos idiomas.
3. Un **decoder autoregresivo** junto con un **vocoder** integrado que genera la onda final.

Entre sus características más importantes destacan su solidez frente a variaciones en la calidad del audio de referencia, la buena reproducción de acentos, una prosodia natural y una menor tendencia a producir artefactos perceptibles. Como desventajas, puede introducir ligeros ruidos si el speaker embedding no está suficientemente limpio y, además, su uso implica un mayor consumo de recursos computacionales en comparación con modelos más ligeros.

3.2. YourTTS

YourTTS es un modelo que sigue un enfoque **zero-shot**, lo que permite sintetizar la voz de un audio sin necesidad de entrenamiento adicional. Su diseño combina varios módulos coordinados, y está compuesto por un speaker encoder basado en **Resemblyzer** para generar la huella vocal, un modelo generativo con **VAE** que captura variaciones naturales en el timbre y la prosodia, y un **vocoder HiFi-GAN**, que es el encargado de transformar las representaciones acústicas en audio final.

Por otra parte, este modelo destaca por su rapidez y ligereza, y suele reproducir el timbre con bastante fidelidad siempre que el audio de referencia esté limpio. No obstante, presenta ciertas limitaciones ya que ofrece un control prosódico más reducido, y puede generar vibraciones o inestabilidades en frases largas resultando más sensible al ruido y a la presencia de silencios en la muestra de voz utilizada como referencia.

4. Pipeline de Implementación

4.1. Entorno y Docker

Con el objetivo de garantizar la reproducibilidad y evitar problemas de configuraciones locales, se ha desarrollado un entorno completamente aislado mediante Docker. El contenedor incluye:

- Python 3.10.
- Dependencias específicas de coqui-TTS y modelos XTTS / YourTTS.
- Librerías para la extracción de embeddings como Resemblyzer y SpeechBrain.
- FFmpeg para la gestión y la conversión de audio.
- Build-essential, necesario para compilar paquetes como webrtcvad.

Tras varias iteraciones y resolución de incompatibilidades entre versiones de *numpy*, *transformers* y componentes del ecosistema TTS, se ha conseguido un contenedor estable y capaz de ejecutar sin errores tanto la generación de audios como la evaluación de similitud.

4.2. Estructura del proyecto

La organización del proyecto es modular con el fin de facilitar pruebas, depuración y extensiones posteriores:

- /src: contiene toda la lógica de inferencia, evaluación y utilidades.
- /audio: almacena el audio de referencia y las voces sintetizadas por cada modelo.
- /tmp_spkrec: es creada automáticamente por el speaker encoder SpeechBrain para almacenar los vectores temporales del speaker encoder.

4.3. Scripts principales

- **generar_xtts.py:** Ejecuta el proceso de síntesis mediante XTTS cargando el speaker embedding, generando el audio y guardándolo en la carpeta /audio.
- **generar_yourtts.py:** Sigue el mismo flujo que el anterior archivo, pero utilizando el modelo YourTTS, permitiendo así comparar ambas aproximaciones.
- **evaluar.py**: Extrae embeddings de los tres audios (original, XTTS y YourTTS) y calcula sus similitudes mediante cosine similarity, lo que permite cuantificar la proximidad entre la voz generada y la voz real.

5. Resultados

Tras el procesamiento del audio original y la ejecución de ambos modelos, se obtuvieron tres archivos finales:

1. **Audio original recortado**
2. **Audio sintético generado con XTTS**
3. **Audio sintético generado con YourTTS**

```
 CARGANDO AUDIOS
Origen: audio/audio_original.wav
XTTS: audio/output_xtts.wav
YourTTS: audio/output_yourtts.wav
```

A continuación, se muestra el análisis comparativo realizado según criterios como naturalidad, similitud vocal y estabilidad.

5.1. Naturalidad

La evaluación auditiva muestra diferencias claras entre los modelos:

Modelo	Naturalidad	Observaciones
XTTS	Alta	Prosodia fluida, pausas bien colocadas y ritmo estable
YourTTS	Media	Sonido más robótico, vibraciones ocasionales y entonación irregular

En general, el modelo XTTS ofrece una naturalidad superior, especialmente en frases largas donde mantiene mejor la estabilidad y la expresividad, mientras que YourTTS muestra una tendencia a un tono más monótono y menos controlado.

5.2. Similitud vocal

Se calcularon dos embeddings independientes para estimar la similitud de timbre y características del hablante.

- **Métrica 1: ECAPA-TDNN (SpeechBrain)**

Esta métrica analiza la identidad vocal de forma robusta:

Comparación	Similitud ECAPA
Original vs XTTS	0.6681
Original vs YourTTS	0.3017

```
ECAPA – Similitud Origen vs XTTS:     0.6681
ECAPA – Similitud Origen vs YourTTS: 0.3017
```

- El modelo XTTS obtiene un valor notablemente más alto, lo que confirma que captura mejor la identidad del hablante.
- El modelo YourTTS muestra una similitud baja debido a variaciones prosódicas y cierta inestabilidad en la formación del timbre.

- **Métrica 2: Resemblyzer (Cosine Similarity)**

Este encoder está optimizado para capturar características finas de la voz:

Comparación	Similitud Resemblyzer
Original vs XTTS	0.8951
Original vs YourTTS	0.7177

```
MÉTRICA 2: Resemblyzer (Cosine Similarity)
Loaded the voice encoder model on cpu in 0.02 seconds.
Resemblyzer – Similitud Origen vs XTTS:     0.8951
Resemblyzer – Similitud Origen vs YourTTS: 0.7177
```

- El modelo XTTS vuelve a ser el más similar, con un valor muy alto (≈0.90).
- El modelo YourTTS logra una similitud moderada, suficiente para reconocer rasgos del timbre, aunque menos consistente.

En conclusión, XTTS supera a YourTTS en ambas métricas, mostrando un timbre más cercano al original y mayor estabilidad.

5.3. Estabilidad y artefactos

En cuanto a XTTS, mantiene una señal limpia, sin clics ni vibración audible, además de que su forma de onda es estable incluso en transiciones rápidas. Por otro lado, YourTTS tiene presencia de ligeros temblores en consonantes continuas y una pequeña fluctuación en el volumen a lo largo de la frase. Estos resultados coinciden con las métricas, en las cuales se concluye que XTTS produce un sonido más uniforme y menos propenso a artefactos.

5.4. Conclusión de comparaciones

Tanto las métricas cuantitativas como la percepción humana coinciden en que el modelo **XTTS** ofrece una mayor similitud con el original, mejor prosodia y menos artefactos, mientras que **YourTTS**, aunque sea más ligero, produce una voz algo menos estable y con menor naturalidad.

6. Discusión Crítica

Los resultados obtenidos confirman que ambos modelos son capaces de realizar clonación de voz en modo *zero-shot*, aunque su rendimiento varía significativamente según las condiciones del audio de entrada y la complejidad prosódica del texto.

El modelo XTTS destaca por su estabilidad y naturalidad ya que mantiene el ritmo, la entonación y el estilo del audio de forma más consistente, y genera una voz más cercana a la original. Esta robustez se aprecia especialmente en fragmentos breves, donde el modelo es capaz de extraer un embedding fiable incluso si el audio contiene pequeñas variaciones.

Por otra parte, el modelo YourTTS resulta más sensible a ruidos, pausas o variaciones en la muestra inicial. Aunque es un modelo rápido y ligero, su prosodia tiende a ser más uniforme y menos expresiva, lo que en ocasiones se traduce en un sonido algo más mecánico o con vibraciones en fonemas sostenidos. Aun así, funciona adecuadamente cuando se dispone de un fragmento limpio y estable.

Un punto clave que muestra este análisis es la gran importancia del preprocesamiento del audio. El recorte previo, en el cual se eliminaron silencios y segmentos contaminados, tuvo un impacto en la similitud de timbre y en la calidad final de la síntesis, ya que influyó, en algún caso, más que la elección del modelo. Esto refuerza que la calidad del embedding del audio puede ser tan importante como la arquitectura del TTS.

7. Conclusiones

El preprocesamiento del audio, incluyendo el recorte, la normalización y el ajuste del formato, es un paso clave para obtener una clonación de voz precisa y estable. A partir del análisis, se concluye que el modelo XTTS destaca como el más equilibrado y fiable, combinando naturalidad, estabilidad temporal y una alta fidelidad al timbre del hablante. Las métricas objetivas ECAPA-TDNN y Resemblyzer refuerzan esta idea, ya que sus resultados están por encima del modelo YourTTS en cuanto a similitud vocal.

Por otra parte, el modelo YourTTS es una alternativa eficiente y rápida, especialmente útil en contextos con recursos limitados, ya que presenta menor control prosódico y una mayor sensibilidad a ruido o a silencios en la muestra de referencia.

En resumen, el pipeline implementado, basado en Docker, scripts de síntesis y procedimientos de evaluación, garantiza reproducibilidad y facilita la ejecución de la práctica en cualquier entorno. En conjunto, este análisis permite comprender con claridad todo el flujo de un sistema moderno de TTS, analizando desde la preparación del audio hasta la comparación cualitativa y cuantitativa entre modelos.

Enlace al repositorio: *https://github.com/carlotadelriego/Practica4_SistemasInteractivos.git*