

La voz del agente: qué es el TTS

Datos de esta lección

Fase	F1 — Cimientos, seguridad y arquitectura del sistema
Módulo	M1.2 — Modelos mentales: radiografía de un agente telefónico
Puntos de profesionalidad al superar el test	15

Objetivo práctico

Al terminar esta lección podrás explicar qué es el TTS (síntesis de voz), qué papel cumple en un agente telefónico y qué factores influyen en que una voz generada suene natural o artificial.

Resultado que obtendrás

Un modelo mental claro de la última pieza del flujo de una llamada: cómo el texto generado por el LLM se convierte en un audio que la persona que llama escucha.

Dónde encaja dentro del sistema final

En el sistema TELÉFONO → AGENTE → CEREBRO → VOZ → AUTOMATIZACIÓN → CALENDARIO, el TTS es la «boca»: convierte el texto que ha decidido el LLM en audio hablado, justo antes de que ese audio llegue de vuelta a la persona por teléfono.

Conocimientos previos

- Lecciones anteriores: LLM (cerebro) y STT (oído).

Herramientas necesarias

- Un navegador con NVDA activo, en modo de exploración.

Posible coste

Esta lección es conceptual. El coste real del TTS se explica con cifras concretas en el módulo 2.1 y se practica en el módulo 2.2 con ElevenLabs.

Modelo mental

Piensa en las cuerdas vocales del telefonista: una vez que el cerebro (LLM) ha decidido qué decir, hace falta un mecanismo que lo pronuncie en voz alta, con una entonación natural. Eso es exactamente lo que hace el TTS.

Explicación

TTS son las siglas de *Text To Speech* (texto a voz). Convierte el texto generado por el LLM en un archivo de audio con voz natural, que se reproduce a la persona que llama. La calidad de un TTS moderno se mide en varios ejes: la naturalidad de la entonación, la latencia (cuánto tarda en generar el audio desde que recibe el texto), la claridad de la pronunciación, y la posibilidad de elegir o incluso clonar una voz concreta. El módulo 2.2 de este curso trabaja en profundidad con ElevenLabs, una de las plataformas de referencia en generación de voz en español, incluyendo parámetros como estabilidad y similitud que afinan cómo suena la voz resultante.

Una latencia alta en el TTS —el tiempo que tarda en generar el audio— es una de las causas más frecuentes de que una conversación telefónica con un agente de IA se sienta «robótica» o con silencios incómodos, incluso cuando el LLM ha respondido con inteligencia y rapidez.

Vocabulario nuevo

TTS

Piénsalo así: Las cuerdas vocales del telefonista.

Text To Speech: tecnología que convierte texto escrito en audio de voz natural.

Ejemplo: El texto «Su cita ha quedado confirmada para el jueves a las diez» se convierte, mediante TTS, en el audio que escucha la persona que llama.

Latencia

Piénsalo así: El tiempo que tarda alguien en empezar a responder después de que le hagan una pregunta.

Tiempo que transcurre entre que un sistema recibe una entrada y produce su respuesta. En telefonía con IA, la suma de las latencias de STT, LLM y TTS determina cuánto silencio percibe la persona que llama antes de oír una respuesta.

Ejemplo: Una latencia de TTS demasiado alta produce una pausa incómoda antes de que el agente empiece a hablar.

Preparación

No se requiere preparación técnica.

Procedimiento paso a paso

1. **Contexto:** Vas a localizar los dos términos nuevos de esta lección, «TTS» y «Latencia», en el Glosario acumulativo de la plataforma.

Acción de teclado: Abre el Glosario (/glosario.php) y pulsa **H** hasta el encabezado de nivel 3 «TTS», dentro del bloque «T».

Respuesta esperada de NVDA: NVDA debería anunciar «TTS, encabezado de nivel 3».

Qué significa: Has confirmado que el término está documentado y es accesible por navegación de encabezados.

Acción siguiente: Lee la definición y el ejemplo con la flecha abajo, y sigue pulsando H hasta llegar al encabezado de nivel 3 «Latencia», también en el bloque «L».

2. **Contexto:** Comprueba que puedes volver directamente a la lista de letras del glosario sin recorrer toda la página hacia atrás.

Acción de teclado: Pulsa **Inicio** (Home) para volver al principio de la página, o pulsa **D** (marca de referencia/landmark) para saltar a la región de navegación «Letras del glosario».

Respuesta esperada de NVDA: NVDA debería anunciar «Letras del glosario, navegación» o el nombre equivalente de la región.

Qué significa: La navegación por regiones (landmarks) te permite saltar directamente a bloques funcionales de una página larga sin leerla entera.

Acción siguiente: Desde ahí, pulsa Tab para recorrer los enlaces de letras y salta a otra que te interese, si quieres seguir explorando.

Errores frecuentes y diagnóstico

La conversación se siente lenta o con silencios largos antes de cada respuesta del agente.

Cómo localizarlo: Puede deberse a la latencia acumulada de STT + LLM + TTS, no necesariamente a un fallo, sino a una configuración mejorable.

Solución: Se revisa en el módulo 2.3, donde se ajustan parámetros de latencia al ensamblar el agente en Vapi o Retell.

Comprobación final

Explica con tus palabras la diferencia entre el LLM (decide qué decir) y el TTS (lo convierte en voz). Si puedes explicar por qué una latencia alta en el TTS afecta a la naturalidad de la llamada aunque el LLM funcione perfectamente, la comprobación es positiva.

Qué acabas de conseguir

Completas el modelo mental de las tres piezas centrales del cerebro-oído-boca del agente: STT, LLM y TTS, la base para entender el resto del sistema.

Ejercicio práctico

Escucha con atención cualquier sistema de voz automatizado que ya exista (un contestador, un asistente de voz de tu teléfono) y anota si notas algún «silencio» antes de que empiece a responder. Ese silencio es, casi siempre, latencia acumulada de las piezas del sistema.

Resumen de lo imprescindible

- TTS significa Text To Speech: convierte texto en audio de voz.
- Es la última pieza del flujo antes de que la persona escuche la respuesta.
- La latencia del TTS (y la de todo el sistema) afecta directamente a la naturalidad percibida de la llamada.
- ElevenLabs es la plataforma de referencia de este curso para TTS en español (módulo 2.2).

No necesitas aprender todavía

Todavía no necesitas configurar ninguna voz real: eso se practica paso a paso en el módulo 2.2.

Estoy preparado para continuar si puedo...

- Explicar qué es el TTS y diferenciarlo del LLM.
- Explicar qué es la latencia y por qué afecta a la naturalidad de una llamada.
- Ubicar correctamente STT, LLM y TTS en el flujo de una llamada.

Test de comprobación

Para registrar el resultado y obtener Puntos de profesionalidad, realiza este test desde la versión web de la lección. Las opciones se muestran aquí sin señalar la respuesta correcta.

1. El LLM de un agente genera una respuesta correcta e inteligente en menos de un segundo, pero la persona que llama percibe casi dos segundos de silencio antes de escuchar nada. ¿Qué pieza es más probable que esté causando ese retraso adicional?
 - a. El STT, porque siempre es la pieza más lenta del sistema.
 - b. El TTS, si su latencia de generación de audio es alta, el silencio percibido puede deberse a esa espera adicional aunque el LLM ya haya respondido.
 - c. El calendario, porque cualquier retraso en la conversación viene de la automatización.
2. ¿Qué diferencia principal hay entre el STT y el TTS dentro del flujo de una llamada?
 - a. El STT convierte voz en texto al principio del flujo; el TTS convierte texto en voz al final, justo antes de que la persona escuche la respuesta.
 - b. Son la misma tecnología aplicada dos veces seguidas.
 - c. El TTS ocurre antes que el STT en cada turno de la conversación.
3. Un agente usa una voz de TTS que, según las pruebas realizadas, suena claramente robótica y con una latencia alta. ¿Qué efecto directo es más probable que note la persona que llama, aunque el LLM responda con total precisión?
 - a. No notará ninguna diferencia, porque la calidad del TTS no afecta a la experiencia de la llamada.
 - b. Percibirá silencios incómodos antes de cada respuesta y una entonación

- poco natural, aunque el contenido de la respuesta sea correcto.
- c. La llamada se cortará automáticamente por parte de la operadora telefónica.

Fuentes

- [Text to speech — OpenAI \(consultado el 2026-08-30\)](#)

Última actualización de este contenido: **2026-02-10**.