Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Checklist: traducción de llamadas en tiempo real

Antes de firmar con un proveedor o arrancar la implementación, revisa cada punto.

1. Proveedores por capa

Capa 1 — Voz a texto (STT)
Deepgram
OpenAI Whisper (API)
Google Speech-to-Text
AWS Transcribe Streaming
Azure Cognitive Services

Capa 2 — Traducción
DeepL API
Google Cloud Translation
OpenAI GPT (uso contextual)
AWS Translate

Capa 3 — Texto a voz (TTS)
ElevenLabs
OpenAI TTS
Google Cloud TTS
AWS Polly
Azure Neural Voice

Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Alternativas all-in-one (las 3 capas en un solo proveedor)
OpenAI Realtime API
Twilio Voice + Media Streams
Vonage AI Studio

2. Preguntas técnicas clave

Latencia end-to-end
¿Cuál es la latencia promedio y el P95? ¿Cómo se comporta bajo carga? Por encima de 500ms la conversación se siente artificial.

Protocolo de audio
WebSocket o WebRTC? Que codecs soporta (ulaw, PCM 16-bit, Opus)? El proveedor recibe audio crudo o requiere formato específico?

Streaming vs batch
El STT opera en modo streaming real (token a token) o espera a que termine la frase? La diferencia es 200 ms vs 1.5s de latencia percibida.

Manejo de interrupciones (barge-in)
¿Qué pasa si el usuario habla mientras el sistema reproduce audio? Se corta automáticamente o hay que manejarlo manualmente?

Reconexión automática
El WebSocket tiene lógica de reconnect? ¿Qué pasa con el audio en tránsito si el socket se cae a las 3am?

Soporte de idiomas y dialectos
¿Distingue el español neutro de rioplatense, mexicano, colombiano? ¿Cómo maneja acentos fuertes o ruidos de fondo?

3. Puntos de falla comunes

Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

Silencio detectado como corte Pausas naturales de 1-2 segundos activan el VAD (Voice Activity Detection) y cortan el stream prematuramente. Revisa el threshold configurable.
Acumulacion de latencia entre capas Cada hop agrega latencia: STT + traducción + TTS puede sumar 800 ms-2s si no se paraleliza correctamente. Medí cada capa por separado.
Rate limiting silencioso Los proveedores aplican rate limits por minuto/hora que no siempre generan error explícito. Una llamada puede cortarse sin traza clara en los logs.
Eco y feedback de audio El TTS que se reproduce puede ser capturado por el micrófono y retraducido. Sin AEC (Acoustic Echo Cancellation), el sistema entra en loop.
Falta de observabilidad en producción Sin métricas por segmento (latencia STT, traducción, TTS, errores por capa), diagnosticar una caída a las 3am es prácticamente imposible.
Contexto perdido entre frases Si la traducción no mantiene contexto conversacional, los pronombres y referencias quedan ambiguos. Cada frase traducida como unidad aislada pierde coherencia.

4. Preguntas para el equipo técnico del proveedor

01	¿Tienen SLA de uptime para el endpoint de streaming? ¿Que garantizan en horarios de baja demanda?
02	¿Cual es el timeout máximo de conexión WebSocket antes de que el servidor cierre el socket?
03	¿Dónde está el servidor más cercano a América Latina (o región de interés)?
04	Los logs incluyen latencia por segmento o solo timestamp de entrada/salida?
05	¿Qué pasa si el rate limit se supera durante una llamada activa? ¿Se cae la llamada, se degrada, o hay un buffer?
06	El audio de las llamadas se almacena en sus servidores? ¿Por cuánto tiempo? ¿Bajo que jurisdicción legal?
07	Tienen endpoint de health check en status page pública para monitorear incidentes en tiempo real?

Gonzalo Gomez
AI Specialist in Communication and Automation Systems
gonzalo@ggomez.dev

08	Tienen casos de uso documentados con Twilio, Vonage, o plataformas de contact center? ¿Pueden mostrar una integración real?

5. Costos reales por minuto de llamada traducida

Capa	Rango	Unidad
STT	**$0.006 - $0.024**	por minuto de audio
Traducción	**$0.020**	por millón de caracteres
TTS	**$0.015 - $0.030**	por 1.000 caracteres
Telefonía (Twilio)	**$0.013 - $0.022**	por minuto de llamada
TOTAL estimado	**$0.06 - $0.14**	por minuto de llamada traducida

Depende del proveedor, la región y el volumen contratado. Con soluciones all-in-one como OpenAI Realtime API el rango puede ser diferente.

- Gonza | ggomez.dev