AVANCE 4. MODELOS ALTERNATIVOS

Ángel Efraín Luna Martínez - A01795486
Francisco Salvador Hernández Pérez - A01795486
Iker Bring Anaya - A01795270

1. Introducción

En esta etapa se evaluaron diferentes configuraciones y modelos de reconocimiento de voz (Speech-to-Text, STT) con el fin de determinar la alternativa más adecuada para integrar al chatbot de la empresa.

El análisis se centró en modelos que cumplieran las restricciones operativas establecidas: despliegue en Google Cloud Platform (GCP), compatibilidad con los formatos de audio de los canales Google Chat, WhatsApp y Microsoft Teams.

A partir de los resultados previos del Avance 3, se amplió la investigación hacia seis modelos alternativos, combinando motores de nube y opciones open-source para comparar desempeño, precisión, costo y facilidad de integración.

2. Modelos evaluados

Se probaron seis configuraciones que agrupan tanto servicios comerciales como modelos open-source. Cada uno se ejecutó en entorno controlado con audios reales representativos del dominio (consultas de viajes, envíos y resúmenes).

ID	Modelo	Tipo	Configuración / Variante	Despliegue
M1	**Google Speech-to-Text v2 (Standard)**	SaaS	Modelo base sin adaptación léxica	GCP nativo
M2	**Google Speech-to-Text v2 (Enhanced)**	SaaS	Activación de *Speech Adaptation* (hotwords)	GCP nativo
M3	**Azure Speech to Text (Base)**	SaaS	Configuración estándar multilingüe	GCP (API externa)
M4	**Azure Speech Custom (Dominio)**	SaaS	Entrenado con vocabulario de negocio (shipment, viaje)	GCP (API externa)
M5	**Whisper Small**	Open Source	Preentrenado, CPU	Cloud Run (CPU)

M6	**Whisper Medium**	Open Source	Optimizado, GPU T4	GKE (GPU)

3. Metodología comparativa

Cada modelo se evaluó con un conjunto de audios homogéneo (≈200 notas de voz de 3 a 30 segundos) en español.

Las métricas de comparación fueron las siguientes:

- WER (Word Error Rate): porcentaje de error por palabra.
- CER (Character Error Rate): error por carácter, útil en detección de números.
- Latencia P95: tiempo máximo esperado para procesar el 95 % de los audios.
- Costo aproximado: USD por 1 000 minutos procesados.
- Compatibilidad: formatos admitidos sin conversión (OGG, M4A, WAV).
- Escalabilidad: facilidad de despliegue y monitoreo en GCP.

4. Resultados de la evaluación

Modelo	Tipo	WER (%) ↓	Latencia (s) ↓	Costo aprox (USD / 1k min) ↓	Hardware recomendado	Despliegue sugerido
Google STT v2 Standard	SaaS	13.4	0.95	16–18	No aplica (servicio gestionado)	API REST en GCP

Google STT v2 Enhanced	SaaS	**9.2**	1.10	24–27	No aplica (servicio gestionado)	API REST en GCP
Azure STT Base	SaaS	12.1	0.90	15–20	No aplica (servicio gestionado)	API REST desde GCP
Azure STT Custom	SaaS	10.8	1.30	25–28	No aplica	API REST desde GCP
Whisper Small	Open Source	11.5	3.10	**≈2–3 (CPU)**	VM e2-standard-8 (8 vCPU, 32 GB RAM)	Cloud Run / Compute Engine
Whisper Medium	Open Source	**8.8**	1.80	**≈6–8 (GPU)**	GPU NVIDIA T4 / L4 + 16 GB RAM	GKE o Vertex AI (GPU)

Costo Whisper calculado a partir de precios GCP:

T4 GPU: ≈$0.35/hora → 1000 minutos (16.6 horas de audio procesado / lote) = ~$6.

CPU (8 vCPU): ≈$0.25/hora → 1000 min ≈ $2–3.

Los costos estimados corresponden al procesamiento de 1 000 minutos de audio, calculados según tarifas oficiales de cada servicio o costos de infraestructura en GCP.
En el caso de modelos open-source como Whisper, el costo proviene exclusivamente del uso de hardware (CPU o GPU), no de licencias.
Whisper Medium requiere una GPU NVIDIA T4 o L4 para lograr un desempeño competitivo (1.8 s P95 por clip de 10 s), mientras que las soluciones SaaS no requieren hardware dedicado al ser totalmente gestionadas.

Principales hallazgos:

Los modelos Whisper Medium y Google Enhanced lograron la menor tasa de error y mantuvieron latencias aceptables.

- Whisper Small, aunque económico, duplicó la latencia del resto.
- Azure Custom mejoró respecto al modelo base, pero no alcanzó la precisión de los mejores competidores.

- Google Enhanced fue el más consistente en audios ruidosos y acentos variados.

- Todos los modelos open-source requirieron conversión a WAV PCM 16 kHz, mientras que los servicios cloud aceptaron directamente OGG/M4A.

5. Selección de los dos mejores modelos

Tras el análisis comparativo, se seleccionaron los siguientes como finalistas:

Google Speech-to-Text v2 (Enhanced):

Ventajas: soporte nativo en GCP, manejo directo de OGG/Opus (WhatsApp) y M4A (Google Chat), alta precisión en números y palabras clave.

Limitaciones: costo por uso, dependencia de API comercial.

Whisper Medium:

Ventajas: modelo open-source robusto, control de datos, adaptable a GPU en GKE, bajo WER en entornos con ruido.

Limitaciones: mayor complejidad operativa, latencia moderada en CPU.

Estos dos modelos representan el mejor balance entre precisión, latencia, costo y compatibilidad con la infraestructura de la empresa.

6. Ajustes y configuración

Para los dos modelos finalistas se definieron parámetros específicos de ajuste:

a) Google STT v2 Enhanced

- Uso de Speech Adaptation (hotwords) con términos del dominio (“viaje”, “shipment”, “resumen”).

- Configuración del modelo “video” (mejor para notas largas) y detección automática de idioma (es-MX).

- Límite de audio: 120 s; timeout operativo: 30 s.

- Uso de Speech Adaptation (hotwords): Se proporcionó una lista de términos clave del dominio ("viaje", "shipment", "resumen").

Este es el ajuste más crítico. El modelo base de Google puede fallar en transcribir correctamente jerga o nombres propios. Al proveer *hotwords*, forzamos al modelo a "escuchar" y priorizar estos términos, reduciendo drásticamente el WER en las palabras más importantes para el negocio.

Parámetros operativos: Se estableció un límite de audio de 120 segundos y un *timeout* de 30 segundos.

- Justificación: Estos límites aseguran que la API responda en un tiempo aceptable para una interacción de chatbot y evitan el procesamiento de audios excesivamente largos que no corresponden al caso de uso.

b) Whisper Medium

- Beam size: 5; temperature: 0.2 (prioriza precisión sobre velocidad).
- Batch size: 8; FP16 activado en GPU para reducir latencia.
- Filtro de ruido y VAD (Voice Activity Detection) previo a la inferencia.
- Ambos modelos entregan texto normalizado con puntuación, capitalización y detección de números, integrándose al pipeline del clasificador de intenciones existente.

7. Modelo final recomendado

Tras el análisis comparativo de los seis modelos y el ajuste de hiperparámetros de los dos finalistas, se recomienda la adopción de Whisper Medium (M6) como el modelo individual final para la solución de Speech-to-Text del proyecto.

Esta decisión se fundamenta en los siguientes puntos clave extraídos de la evaluación:

1. Precisión Superior (WER): Whisper Medium obtuvo la tasa de error de palabra (WER) más baja de toda la comparativa, con un 8.8%. Este resultado supera al modelo SaaS más avanzado, Google STT v2 Enhanced, que alcanzó un 9.2%.
2. Costo Operativo a Escala: El factor decisivo es el costo a largo plazo. Desplegado en GKE con una GPU T4, el costo estimado de Whisper Medium es de ≈$6–8 USD por cada 1,000 minutos procesados. Esto representa un ahorro de más del 70% en comparación con los $24–27 USD estimados para Google STT v2 Enhanced.

3. Soberanía y Control de Datos: Al ser un modelo *open-source* desplegado en la infraestructura de GCP propia, Whisper garantiza control total sobre los datos. Los audios (que pueden ser sensibles) no necesitan salir del entorno controlado de la empresa para ser procesados por una API comercial de terceros.

Esta elección implica un compromiso estratégico consciente: se acepta una mayor complejidad operativa (requiriendo la gestión de un clúster de GKE con GPU) y una latencia P95 ligeramente superior (1.80s vs 1.10s de Google).

Se concluye que los beneficios de una precisión superior, un ahorro significativo a escala y la soberanía total de los datos superan los desafíos de la latencia y la gestión de la infraestructura.

8. Conclusiones

La comparación demostró que los modelos de transcripción basados en redes neuronales profundas ofrecen resultados significativamente superiores a soluciones tradicionales.
Los motores evaluados mostraron diferencias principalmente en costo, velocidad y nivel de personalización.
Con la selección de Whisper Medium como modelo base y Google STT v2 Enhanced como alternativa comercial, el proyecto cuenta con una arquitectura flexible, escalable y alineada con las restricciones de GCP.

El siguiente paso será documentar los requerimientos técnicos para el despliegue controlado (Cloud Run o GKE), establecer métricas de monitoreo y definir las pruebas de validación con audios reales del entorno productivo.

Las pruebas técnicas se realizaron con la siguiente APP:

https://github.com/A01795486/Tracky-STT

Referencias

- Google Cloud. (2023). Introducing Google Cloud Speech-to-Text v2 API. Google Cloud Blog. https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-speech-to-text-v2-api
- Microsoft. (2025). Speech to text documentation — Tutorials & Reference. Microsoft Learn. Recuperado el 7 de octubre de 2025, de https://learn.microsoft.com/en-us/azure/ai-services/speech-service/index-speech-to-text

- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv. https://doi.org/10.48550/arXiv.2212.04356
- Xu, B., Tao, C., Feng, Z., Raqui, Y., & Ranwez, S. (2021). A Benchmarking on Cloud-based Speech-to-Text Services for French Speech and Background Noise Effect. arXiv. https://doi.org/10.48550/arXiv.2105.03409