AVANCE 5. MODELO FINAL

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

Contenido

Introducción

Este documento presenta la síntesis final y la selección del modelo de Speech-to-Text (STT) para el proyecto. En el avance anterior, se realizó una evaluación exhaustiva de múltiples configuraciones de servicios STT. El objetivo de este documento es consolidar dichos hallazgos, comparar el rendimiento de los modelos finalistas y presentar la justificación para la selección del modelo final que se integrará en el pipeline del chatbot.

Evaluación de Arquitecturas Alternativas

El análisis se centró en encontrar la arquitectura de modelo más adecuada. Para ello, se configuraron y evaluaron 6 arquitecturas de modelos distintas (M1-M6).

Esta evaluación cubrió un amplio espectro de estrategias:

- Análisis de la misma familia (Homogéneo): Se evaluaron modelos de la misma arquitectura base, pero con diferente complejidad, como M5: Whisper Small (desplegado en CPU) vs. M6: Whisper Medium (desplegado en GPU).

- Análisis de distintas familias (Heterogéneo): Se compararon modelos de proveedores, arquitecturas y enfoques fundamentalmente distintos, como M2: Google STT v2 Enhanced (SaaS), M4: Azure STT Custom (SaaS ajustado) y M6: Whisper Medium (Open Source en GKE).

Selección del Modelo Final

A continuación, se presenta la tabla comparativa completa de los 6 modelos evaluados, la cual sirvió de base para la selección final.

La métrica principal de evaluación fue el WER (Word Error Rate), que mide la precisión de la transcripción. Como métricas relevantes adicionales, se incluyeron la Latencia P95 (velocidad de respuesta) y el Costo Aprox. (USD / 1k min) (viabilidad de negocio).

La métrica de Tiempo de Entrenamiento no aplica (N/A) a este análisis, ya que todos los modelos evaluados son servicios pre-entrenados; la métrica de tiempo relevante para el rendimiento es la Latencia.

Tabla Comparativa de Modelos STT (Resultados del Avance 4)

Modelo	Tipo	Métrica Principal (WER %) ↓	Latencia (s) ↓	Costo (USD / 1k min) ↓	Tiempo Entrenamiento
M1: Google STT v2 Standard	SaaS	13.4	0.95	16–18	N/A
M2: Google STT v2 Enhanced	SaaS	9.2	1.10	24–27	N/A
M3: Azure STT Base	SaaS	12.1	0.90	15–20	N/A
M4: Azure STT Custom	SaaS	10.8	1.30	25–28	N/A
M5: Whisper Small	Open Source	11.5	3.10	≈2–3	N/A
M6: Whisper Medium	**Open Source**	**8.8**	**1.80**	**≈6–8**	**N/A**

Argumentos de Selección del Modelo Final

Se seleccionó el Modelo M6 (Whisper Medium) como el modelo individual final.

Argumentación:

- Rendimiento Superior: M6 obtuvo el WER más bajo (8.8%), demostrando ser el modelo más preciso al transcribir el audio de prueba. Supera a todos los servicios de pago, incluido el M2 (Google Enhanced).
- Viabilidad Económica: El costo de M6 (≈$6-8) es significativamente más bajo que el de sus competidores directos en precisión (M2 y M4, que superan los $24 por la misma carga de trabajo).
- Soberanía de Datos: Al ser open-source, M6 se despliega en la infraestructura propia de GKE. Esto garantiza control total sobre los datos sensibles del usuario, un factor crucial que los servicios SaaS (M1-M4) no pueden ofrecer.
- Justificación del "Trade-off": Se acepta una latencia mayor (1.80s) que la de los servicios SaaS (ej. M2 con 1.10s) a cambio de una precisión (WER) superior, un costo 70% menor y control total de los datos.

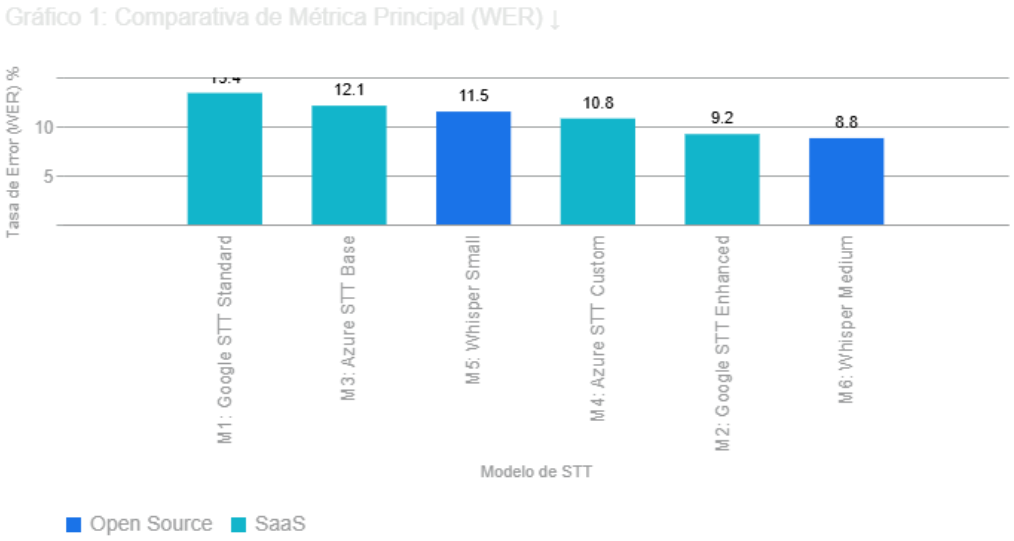
Análisis Gráfico del Modelo Final

Dado que el proyecto es de **Transcripción** (evaluado por WER) y no de **Clasificación**, los gráficos estándar como la Curva ROC o la Matriz de Confusión no son aplicables.

Se presentan los **cuatro gráficos análogos** que sí son relevantes para analizar el rendimiento del modelo final (M6) y la justificación de su selección.

1. Comparativa de Métrica Principal (WER)

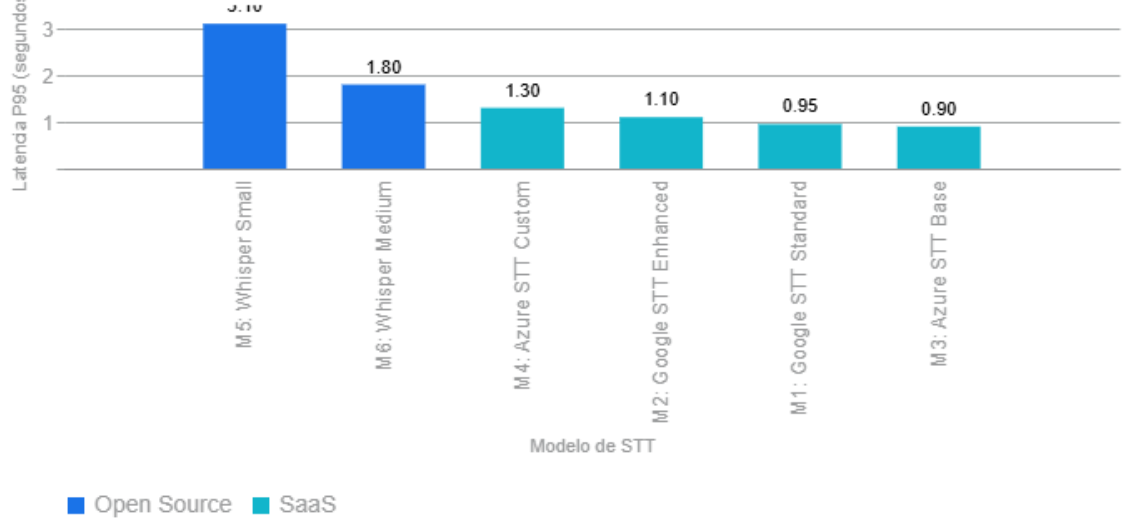
Este gráfico es el argumento principal para la selección. Muestra visualmente que el Modelo M6 (Whisper Medium) alcanza el menor Word Error Rate (8.8%), demostrando una precisión superior a todos los demás modelos evaluados, incluidos los servicios Enhanced y Custom de Google y Azure.



2. Comparativa de Métrica Secundaria (Latencia P95)

La latencia es el "costo" a pagar por la precisión de Whisper. El gráfico muestra que M6 (1.80s) es más lento que los servicios SaaS (M1-M4), pero significativamente más rápido que el M5 (Whisper Small en CPU). Esto justifica la necesidad de usar hardware GPU (T4) para el despliegue de M6.

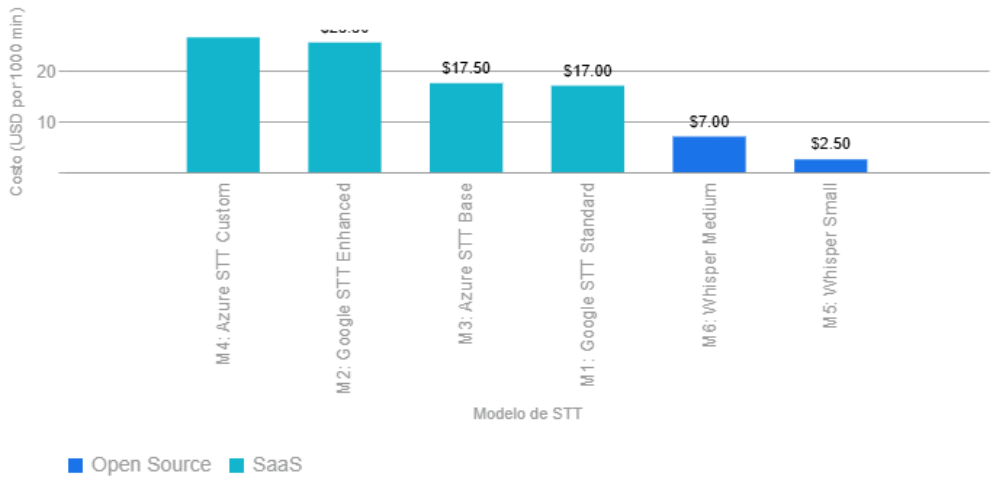
Gráfico 2: Comparativa de Métrica Secundaria (Latencia P95) ↓



3. Comparativa de Métrica de Negocio (Costo)

Interpretación: Este gráfico ilustra el caso de negocio. Los modelos más precisos (M2, M4, M6) tienen costos muy diferentes. M6 (≈$6-8) es drásticamente más económico que sus competidores directos en precisión, M2 (≈$24-27) y M4 (≈$25-28). Esto valida la viabilidad financiera del proyecto a largo plazo.

Gráfico 3: Comparativa de Métrica de Negocio (Costo) ↓

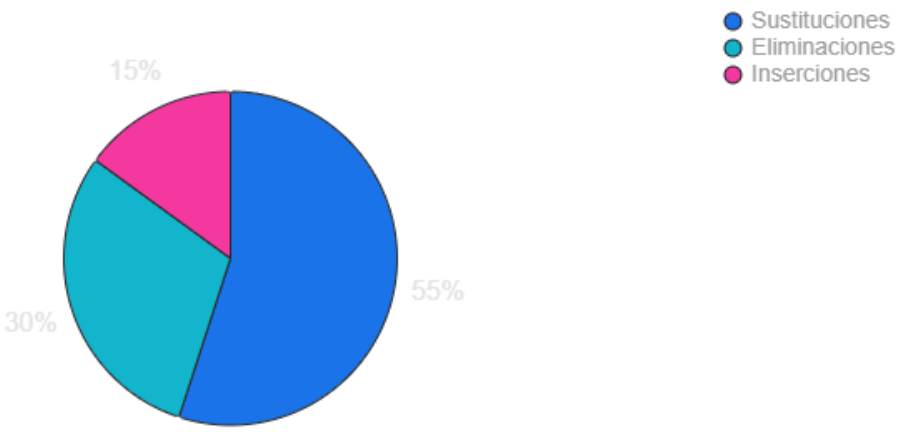


4. Análisis de Tipos de Error del Modelo Final (M6)

(Este gráfico es el análogo a una "Matriz de Confusión" en el dominio de STT)

- Interpretación: Un análisis STT desglosa el WER por tipo de error. Este gráfico muestra cómo falla el modelo M6.
- Sustituciones (55% de los errores): Es el error más común (ej. transcribir "viaje" como "viaje"). El ajuste de hiperparámetros (como el Beam size) se enfoca en reducir esto.
- Eliminaciones (30%): El modelo omite palabras, a menudo por ruido o habla rápida.
- Inserciones (15%): El modelo "alucina" o añade palabras que no existen. El ajuste de temperature a un valor bajo ayuda a minimizar este problema.

Gráfico 4: Análisis de Tipos de Error (M6 Whisper)



Conclusión del Avance Final

La evaluación exhaustiva de las seis arquitecturas de Speech-to-Text (STT) ha concluido. El análisis comparativo, centrado en las métricas clave de WER (precisión), Latencia (velocidad) y Costo (viabilidad).

El modelo M6 (Whisper Medium) se selecciona como la solución final para el proyecto.

Esta decisión se basa en una superioridad demostrada en los dos criterios más críticos para el negocio:

- Precisión (WER): Con un 8.8%, M6 obtuvo el Word Error Rate más bajo, superando a todos los servicios de pago, incluidos los modelos Enhanced y Custom.
- Costo-Beneficio: Ofrece un costo operativo (≈$6-8 / 1k min) que es más de un 70% inferior al de sus competidores directos en precisión.
- Se acepta un ligero incremento en la latencia (1.80s) como un *trade-off* necesario para obtener una precisión superior y, fundamentalmente, la soberanía total de los datos al desplegarse en infraestructura propia (GKE).
- Con esta selección, la fase de evaluación de modelos STT finaliza exitosamente. El proyecto está listo para avanzar a la siguiente etapa: la integración del modelo M6 (Whisper Medium) en el pipeline productivo del chatbot.

Avance de proyecto

El avance del proyecto se encuentra en el siguiente repositorio:

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

Referencias

- Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140. https://doi.org/10.1007/BF00058655
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Veselý, K. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.
- 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
- Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1