

INFORME DE AVANCE TÉCNICO

Proyecto: CatchAI – Sistema MVP de traducción de voz a Lengua de Señas Chilena (LSCh)

Fecha: 03/09/2025

Responsable de informe: Equipo Capstone Puerto Montt

# INFORME DE AVANCE – PROYECTO CANAL 13C

## 

## **Resumen del avance**

Durante este hito se presentaron avances en dos líneas principales del proyecto:

* Reconocimiento de voz y transcripción (Luis Gómez)
* Captura de movimiento y animación 3D del avatar (Marcos Bombalas, Marcos Hernández)

## Avances técnicos

### Reconocimiento de voz y transcripción de texto

**Responsable:** Luis Gómez

* Se evaluaron distintos modelos de reconocimiento de voz, particularmente la versión destilada Fast Whisper, priorizando modelos livianos por limitaciones de hardware (sin soporte CUDA).
* Se logró implementar un flujo funcional con CPU que permite capturar voz mediante micrófono, transcribirla en tiempo real (aproximadamente 0.5 segundos de procesamiento por bloque de 2 segundos).
* Se realizaron pruebas con frases como “Hola, ¿cómo estás tú?”, obteniendo buenos resultados con los modelos pequeños y medianos.
* Se generó un archivo transcripciones.txt para almacenar las pruebas de voz a texto.
* El modelo seleccionado se considera óptimo dentro de las restricciones del equipo, siendo funcional para pruebas de MVP.

**Actualización:** Actualmente se continúa mejorando este modelo, incorporando una búsqueda de las palabras transcritas en el Diccionario de Lengua de Señas Chilena (LSCh).

* Para las palabras no registradas, se utiliza el deletreo como solución temporal.
* En futuras iteraciones se proyecta entrenar un modelo con fine tuning que permita identificar expresiones, agrupar palabras similares por significado y asignar la seña correspondiente.

### 2. Captura de movimiento y animación 3D

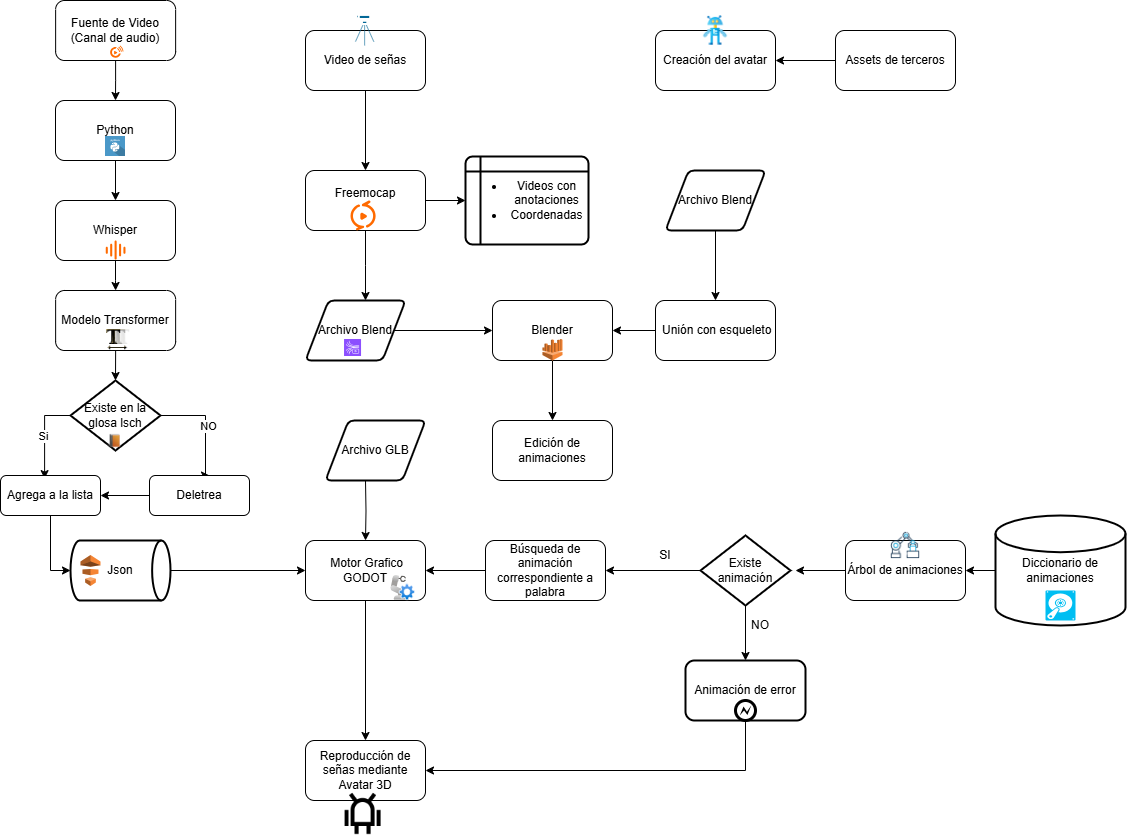
**Responsables:** Marcos Bombalas, Marcos Hernández

* Se implementó el sistema FreeMoCap para capturar señas en formato de coordenadas tridimensionales utilizando la IA MediaPipe, las cuales estarán disponibles, pero no utilizables ya que se utilizarán animaciones de blender.
* Se calibraron múltiples cámaras mediante el uso de una tabla Charuco, generando archivos de calibración TOML
* Las tomas aún se encuentran en fase de mejora y ajuste de calibración, ya que el ambiente de captura aún no ha sido completamente montado debido a la falta de equipos definitivos.
* Actualmente se utilizan equipos de baja gama, sin GPU y webcams de distintas características, lo que afecta la precisión y calidad de la captura.
* Se transformaron los datos en animaciones dentro de Blender, distinguiendo esqueleto y mesh.
* Se desarrolló un script en Python para transferir animaciones al esqueleto principal, exportando el resultado como archivo .GLB.
* El archivo .GLB fue correctamente importado al motor gráfico Godot, permitiendo reproducir animaciones en el entorno.
* Se probó un avatar 3D de baja resolución como prueba de integración; la animación fue correctamente ejecutada.
* Se creó la escena 3D y la estructura de carpetas necesarias para almacenar la librería de animaciones.
* Se proyecta reemplazar el avatar por uno de mayor calidad y mejorar la transición entre animaciones.
* **Avatar:** El modelo propuesto por Marcos Hernández es de baja complejidad y será reemplazado por uno de mediana complejidad para pruebas avanzadas. Para el modelo final se contempla un avatar de género femenino, con etnia chilena (facciones locales), pocos accesorios y vestimenta semiformal.
* **Relevancia del hardware:** Se señala que tanto para la animación como para el procesamiento general del sistema, los resultados y tiempos de ejecución mejorarán considerablemente con mejores equipos (equipos con GPU y estaciones de trabajo más robustas).

### 3. Diagrama de Flujo del Sistema (propuesta presentada)

El diagrama de flujo del sistema presentado incluye las siguientes etapas:

1. Captura de voz mediante micrófono.
2. Transcripción a texto utilizando modelo Fast Whisper.
3. Búsqueda de la palabra transcrita en el diccionario de señas chileno.  
   * Si existe: se busca en el árbol de animaciones y se reproduce la seña.
   * Si no existe: se procede al deletreo letra por letra.
4. En próximas versiones: agrupación semántica mediante fine tuning para asignar señas por equivalencia.
5. La animación seleccionada se reproduce en el motor gráfico Godot con el avatar 3D.



## Observaciones

* Las pruebas de reconocimiento de voz muestran latencia aceptable, aunque la calidad podría mejorar con una GPU compatible con CUDA.
* La primera integración entre FreeMoCap y Blender fue exitosa; se logró importar animaciones en Godot.
* El avatar de baja resolución permitió verificar la compatibilidad, pero deberá reemplazarse por un modelo definitivo más detallado.
* El flujo completo comienza a consolidarse como arquitectura base para el MVP funcional del proyecto.

**Solicitud adicional:** Se plantea la necesidad de contar con profesionales validados en el área de lengua de señas para que al finalizar el proyecto puedan participar en la validación funcional del sistema y emisión de informes de calidad sobre precisión, fidelidad y comprensión de las animaciones generadas.

**Cronograma y ajustes:** Debido a los retrasos en la formalización de convenios y acuerdos con Canal 13, el cronograma del proyecto se ha visto alterado parcialmente. Por lo tanto, se ha decidido priorizar el desarrollo de un MVP funcional centrado en la traducción, con resultados aterrizados en función de los tiempos del semestre. El objetivo es obtener un producto base robusto y funcional, que sirva como punto de partida escalable para futuras iteraciones del proyecto.