UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería


# Diseño e implementación de un módulo de reconocimiento de voz para el control de sistemas robóticos

Protocolo de trabajo de graduación presentado por Oscar Fernando López, estudiante de Ingeniería Mecatrónica

Guatemala,

2025

# Antecedentes

Los sistemas de reconocimiento de voz son tecnologías en constante desarrollo, que forman parte de una tendencia hacia el desarrollo de interfaces naturales, las cuales tienen el propósito de mejorar la interacción entre humanos y máquinas. Aunque tradicionalmente se ha recurrido a métodos visuales o basados en imágenes, la interacción por voz ha ganado importancia debido a su capacidad para ofrecer interacciones rápidas y flexibles. Sus aplicaciones empiezan desde asistentes virtuales, asistencia en cirugías y se pueden ver en entornos de automatización industrial. Los avances buscan incrementar la precisión en la conversión de voz a texto y asimismo, se busca implementar sistemas que se adapten a distintos hablantes sin importar idioma o pronunciación.

Anteriormente, en la Universidad del Valle de Guatemala se inició una búsqueda por encontrar una interfaz natural que permita el control de sistemas robóticos. De esta forma surge la primera línea de investigación, la cual tuvo como objetivo la interpretación de información visual del sistema Brainlab en movimientos robóticos. [1]

En el proyecto desarrollado [1] se logró implementar una herramienta de procesamiento digital de imágenes y reconocimiento óptico de caracteres donde se consiguió la extracción automática de datos visuales necesarios para configurar la posición de un brazo robótico físico. Sin embargo, el trabajo presentó ciertas limitaciones, especialmente en la adaptabilidad del sistema ante variaciones en la calidad visual, como condiciones de iluminación o ángulos no óptimos de captura de pantalla. En consecuencia, se destaca la importancia de desarrollar métodos alternativos al procesamiento visual, capaces de reducir las restricciones operativas.

Por consiguiente, la presente investigación abre una segunda línea de trabajo centrada en el reconocimiento de voz.

## Diseño de un sistema de reconocimiento de voz para un brazo robótico para cirugía laparoscópica

En el trabajo realizado por Ricardo Pastor [2] sobre un sistema de reconocimiento de voz para un brazo robótico para asistencia en cirugías, el autor trabajó en un hardware embebido el cual consiste en una Raspberry Pi, una tarjeta de sonido USB Manhattan, un micrófono, una microSD y un botón físico. Asimismo, Python fue el lenguaje de programación principal ya que posee una gran variedad de librerías oportunas para el manejo de audio.

El principal objetivo de este trabajo era crear un sistema que reconociera comandos de voz y poder transcribirlos internamente e interpretarlos como acciones que se deben ejecutar por parte del robot asistente. Por lo tanto, se propuso un módulo de 9 pasos que pueda pasar de un estímulo auditivo a una acción física.

La primera etapa corresponde a la eliminación de silencio y ruido, cuyo objetivo es detectar la porción activa de la señal de voz y descartar los fragmentos que contienen ruido de fondo o silencio, lo cual reduce el procesamiento innecesario y mejora la precisión del reconocimiento. Posteriormente, se aplica una técnica conocida como preénfasis, que consiste en reforzar las frecuencias altas de la señal, ya que estas contienen información fonética crítica

que podría perderse si se tratara la señal de forma plana.

En la tercera etapa, una vez la señal esté preprocesada, es dividida en pequeñas ventanas; en cada una de estas ventanas se aplica la ventana de Hamming, una función que suaviza los bordes y evita discontinuidades al aplicar la transformada de Fourier. Luego, se utiliza la Transformada de Fourier de Tiempo Reducido (STFT), la cual permite observar cómo varían las frecuencias a lo largo del tiempo, generando un espectrograma de la señal.

Posteriormente, en la penúltima etapa se extraen los coeficientes cepstrales en la frecuencia de Mel (MFCC). Esta técnica convierte la información espectral en un conjunto compacto de coeficientes que representan cada segmento del habla. Finalmente, la última etapa para comparar los comandos hablados con las muestras almacenadas en el sistema se aplica el método de alineamiento temporal dinámico (DTW), el cual permite medir la similitud entre secuencias que pueden diferir en duración o velocidad de pronunciación. [2]

### Aprendizaje por refuerzo de un *parser* semántico óptimo en DRT

El trabajo realizado por Jessenia Piza [3] tuvo un énfasis en el aprendizaje por refuerzo de un *parser* semántico. Su trabajo pretendía sustituir los *parsers* tradicionales, ya que estos requieren de reglas manuales complejas, difíciles de mantener y poco escalables. Por lo que decidió incluir el aprendizaje por refuerzo profundo para que un agente aprenda a generar estas estructuras semánticas de forma autónoma, partiendo de recompensas por cada estructura lógica acertada.

La solución desarrollada por Piza empezaba desde un entorno simulado donde un agente recibe una oración simple y debe ser capaz de convertirla en una estructura DRS (*Deep Reinforcement Learning*). El entorno permite realizar acciones como moverse a lo largo de las palabras y darles una interpretación como sujeto, verbo o sustantivo, y también permite agregar referentes y condiciones a la oración. Es así como se iba recompensando al agente únicamente si respondía una pregunta correctamente donde se demostrara la comprensión lógica de la oración.

Para lograr esto se usó una red Deep Q-Network en donde la entrada es un vector que representa el estado del agente, es decir, las palabras procesadas y la estructura DRS actual, y su salida es la selección entre una de las cinco acciones disponibles.

Dentro del entrenamiento para el modelo, la autora notó mucha inestabilidad por lo que decidió ajustar dos hiperparámetros, los cuales eran Gamma ($\gamma$) y Epsilon ($\epsilon$): Gamma decide cuánta importancia se otorga a las recompensas futuras, y Epsilon es responsable de equilibrar la elección entre acciones conocidas y acciones nuevas. Estos ajustes se realizaron con el propósito de obtener un mejor desempeño, lo cual fue efectivo y lo demostró en una tabla comparando los rendimientos. [3]

## Justificación

La integración de la robótica en diversos sectores ha sido una herramienta fundamental para reducir la carga operativa y facilitar tareas que antes eran complejas o incluso inviables

para el ser humano. Sin embargo, uno de los principales desafíos en la interacción del humano al robot es el control intuitivo y seguro que permitan una comunicación eficiente sin generar distracciones durante la operación del sistema.

En ese sentido, los sistemas de reconocimiento de voz se presentan como una solución viable y altamente prometedora, ya que permite al usuario emitir comandos de forma verbal sin desviar su concentración y sin intervenir físicamente. Por lo tanto, es capaz de llevar a cabo cambios de posición, de orientación o de fuerza que le sean de utilidad para la aplicación pertinente.

En este trabajo se propone el desarrollo de un sistema que interprete comandos de voz emitidos por el usuario para controlar una articulación robótica. Se busca crear una solución universal que pueda adaptarse al control de sistemas robóticos de distintas áreas, tales como entornos médicos, industriales, académicos o de asistencia personal, contribuyendo así a mejorar la interacción entre humanos y robots.

Anteriormente, en la primera línea de investigación la cual se centra en el procesamiento de imágenes y reconocimiento óptico de caracteres, se determinaron ciertas limitaciones y desafíos que se buscan remover al implementar el reconocimiento de comandos por voz. Dentro de las limitaciones más significativas se encuentra la incapacidad que tiene el modelo de hacer la interpretación de una imagen si esta no está centrada de la forma correcta, ángulo correcto e inclusive si no posee el brillo adecuado. Estos aspectos hacen que haya una gran dependencia de la configuración de la pantalla y no es capaz de hacer ninguna interpretación si alguno de los aspectos mencionados no estén en el punto deseado.

Además, también existen limitaciones con el módulo Tesseract el cual extrae el texto de las imágenes ya que depende mucho de la calidad de la imagen, ya que si hay mucho ruido visual entonces su precisión baja considerablemente.

De igual manera, se reconocen limitaciones para esta segunda línea de investigación ya que aspectos como calidad del micrófono y cantidad de ruido ambiental pueden afectar la interpretación de comandos. También se debe considerar que la pronunciación de palabras es única, al igual que la velocidad de habla, por lo que esto supone un gran reto para construir un modelo que sepa adaptarse a las características del habla de cada usuario.

Por lo tanto, se busca la construcción de un sistema de reconocimiento de comandos por voz que sea capaz de funcionar correctamente ante situaciones críticas donde se pueda demostrar su adaptabilidad, precisión y seguridad.

# Objetivos

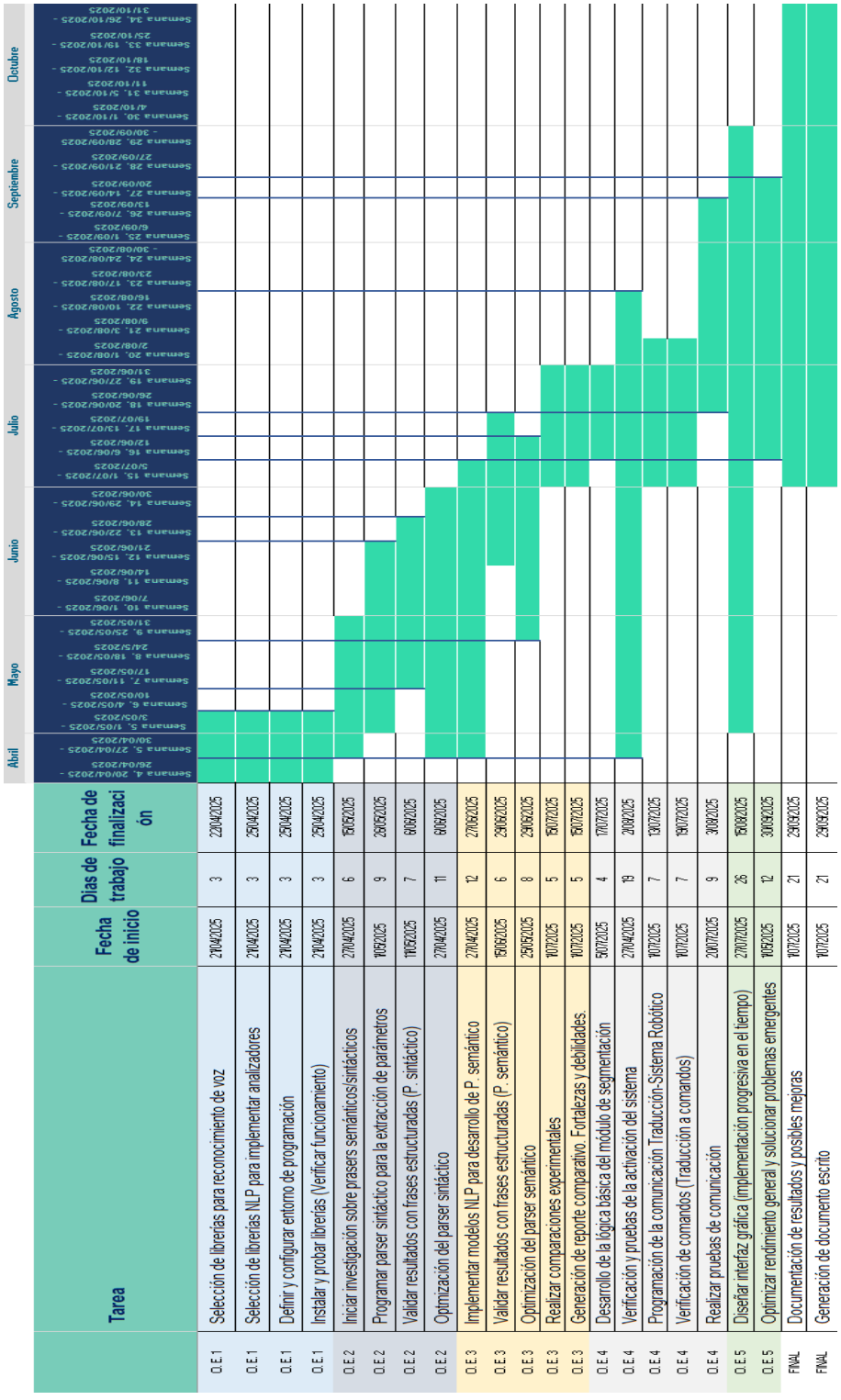
## Objetivo general

Diseñar y desarrollar un sistema de reconocimiento de voz capaz de interpretar comandos para el control de sistemas robóticos.

### Objetivos específicos

- Investigar, evaluar y seleccionar las librerías más adecuadas para la transcripción de voz en tiempo real.
- Desarrollar un analizador sintáctico y un analizador semántico para la interpretación de comandos de voz.
- Validar la efectividad del sistema de interpretación de comandos de voz mediante pruebas con sistemas robóticos disponibles en el Departamento de Ingenieria Electrónica y Mecatrónica de la Universidad del Valle de Guatemala.
- Implementar un módulo de activación por voz que detecte un comando o frase especial para iniciar la comunicación con el sistema robótico y habilitar la recepción de instrucciones.
- Diseñar una interfaz gráfica intuitiva que permita visualizar en tiempo real la transcripción y el análisis de los comandos de voz, así como el estado de ejecución del sistema robótico.

# Cronograma de actividades




| | Tarea | Fecha de inicio | Dias de trabajo | Fecha de finalización |
|---|---|---|---|---|
| O.E.1 | Selección de librerías para reconocimiento de voz | 21/04/2025 | 3 | 22/04/2025 |
| O.E.1 | Selección de librerías NLP para implementar analizadores | 21/04/2025 | 3 | 25/04/2025 |
| O.E.1 | Definir y configurar entorno de programación | 21/04/2025 | 3 | 25/04/2025 |
| O.E.1 | Instalar y probar librerías (Verificar funcionamiento) | 21/04/2025 | 3 | 25/04/2025 |
| O.E.2 | Iniciar investigación sobre prasers semánticos/sintácticos | 27/04/2025 | 6 | 15/05/2025 |
| O.E.2 | Programar parser sintáctico para la extracción de parámetros | 16/05/2025 | 9 | 26/05/2025 |
| O.E.2 | Validar resultados con frases estructuradas (P. sintáctico) | 19/05/2025 | 7 | 60/82/2025 |
| O.E.2 | Optmización del parser sintáctico | 27/04/2025 | 11 | 60/82/2025 |
| O.E.3 | Implementar modelos NLP para desarrollo de P. semántico | 27/04/2025 | 12 | 27/05/2025 |
| O.E.3 | Validar resultados con frases estructuradas (P. semántico) | 15/05/2025 | 6 | 29/05/2025 |
| O.E.3 | Optmización del parser semántico | 25/05/2025 | 8 | 29/05/2025 |
| O.E.3 | Realizar comparaciones experimentales | 16/07/2025 | 5 | 15/07/2025 |
| O.E.3 | Generación de reporte comparativo. Fortalezas y debilidades. | 16/07/2025 | 5 | 15/07/2025 |
| O.E.4 | Desarrollo de la lógica básica del módulo de segmentación | 15/07/2025 | 4 | 17/07/2025 |
| O.E.4 | Verificación y pruebas de la activación del sistema | 27/04/2025 | 19 | 20/82/2025 |
| O.E.4 | Programación de la comunicación Traducción-Sistema Robótico | 16/07/2025 | 7 | 13/07/2025 |
| O.E.4 | Verificación de comandos (Traducción a comandos) | 16/07/2025 | 7 | 19/07/2025 |
| O.E.4 | Realizar pruebas de comunicación | 20/07/2025 | 9 | 30/82/2025 |
| O.E.5 | Diseñar interfaz gráfica (implementación progresiva en el tiempo) | 27/07/2025 | 26 | 15/08/2025 |
| O.E.5 | Optimizar rendimiento general y solucionar problemas emergentes | 16/82/2025 | 12 | 30/08/2025 |
| FINAL | Documentación de resultados y posibles mejoras | 16/07/2025 | 21 | 29/08/2025 |
| FINAL | Generación de documento escrito | 16/07/2025 | 21 | 29/08/2025 |

Figura 1: Cronograma de actividades del proyecto.

## Referencias


[1] S. Boch, «Optimización de la herramienta de procesamiento de imágenes para el sistema Brainlab de Humana, Fase IV,» Trabajo de graduación de licenciatura, Universidad del Valle de Guatemala, 2024, págs. 0-84.

[2] R. Hernández, «Facultad de ciencias de la computación,» Trabajo de graduación de licenciatura, Benémerita Universidad Autónoma de Puebla, ago. de 2018. dirección: https://hdl.handle.net/20.500.12371/7914.

[3] J. Piza, «Parser semántico basado en redes neuronales de tipo Deep Q-Network,» ago. de 2024. DOI: https://doi.org/10.48713/10336_43268.