BENÉMERITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Diseño de un sistema de reconocimiento de voz

para un brazo robótico para cirugía laparoscópica

TESIS

Para obtener el Título

de Licenciado en Ingeniería en Ciencias

de la Computación

Que presenta el C.

RICARDO PASTOR HERNÁNDEZ

ASESORES:

DR. JOSÉ LUIS HERNÁNDEZ AMECA

DRA. ELSA CHAVIRA MARTÍNEZ

Puebla, Pue. Octubre 2018

Agradecimientos:

A mis padres por ser mis superhéroes de la vida real, por apoyarme incondicionalmente durante toda mi vida, desde lo más profundo de mi corazón les agradezco que sigan a mi lado demostrándome día a día su inmenso amor.

A mi hermana, por enseñarme que no es la calificación lo que realmente importa, sino los conocimientos que adquieres en el proceso.

Al amor de mi vida, por apoyarme, por nunca dejar de creer en mí, y por haberme soportado todo este tiempo.

A mis asesores por nunca dejar de creer en la culminación de este trabajo.

A todos y cada uno de mis amigos que, de diferentes maneras, han sido parte de este camino.

Contenido

Introducción

El presente trabajo es parte de un Proyecto general que consiste en el diseño y desarrollo de un brazo robótico para cirugía laparoscópica controlado por voz, la idea fue planteada por el Dr. Alejandro Pedroza Meléndez, académico emérito de la Academia Mexicana de Cirugía. Este trabajo se desarrolla en cuatro fases:

La primera fase fue el diseño y construcción de micro herramientas para cirugía laparoscópica, tesis que fue desarrollada por el hoy Ingeniero en Mecatrónica, Salvador Olivares Hoyos.

La segunda fase es el diseño, desarrollo y construcción de un brazo robótico para cirugía laparoscópica, tesis que se está desarrollando por la estudiante en Ingeniería en Ciencias de la Computación, Valeria Temozihui Tlahuel.

La tercera fase es un sistema de control por voz para poder manipular el brazo robótico para cirugía laparoscópica, tesis presentada a continuación.

La cuarta fase es la integración de las etapas anteriores en el Proyecto general del robot cirujano que consta de un brazo robótico para poder manipular las micro herramientas y el laparoscopio, por medio de un sistema de control que funciona mediante comandos de voz.

Acorde a la tercera fase, en el presente trabajo se desarrolla una propuesta de un sistema de control por voz para un brazo robótico para cirugía laparoscópica.

El presente trabajo se refiere al tema de reconocimiento de voz, en el cual por medio de transductores que transforman señales acústicas en señales eléctricas para ser tratadas por un dispositivo electrónico el cual se encargará de

procesar y comparar estas señales con alguna de las previamente grabadas, en caso de que ambas señales coincidan, el sistema realizará la acción correspondiente.

La presente tesis se basa en conceptos como el Procesamiento Digital de Señales (PDS; Digital Signal Processing o DSP) que consiste en el tratamiento y manipulación de una o varias señales que contienen información con la finalidad de mejorarlas y optimizarlas, estas señales se generan de forma análoga (es decir que son continuas en el tiempo), mientras que un sistema digital maneja información discontinua en el tiempo, por dicha razón la mayoría de los sistemas PDS´s tienen como primera etapa convertir la señal analógica a digital por medio de un CAD, como se muestra en el diagrama a bloques de la figura 1.

Fig. 1. Diagrama a bloques de un ADC

El Reconocimiento Automático del habla (RAH; Automatic Speech Recognition, ASR) es un campo de la computación que tiene como objetivo la comunicación humano – computadora, siendo esto posible cuando un sistema digital recoge una señal de voz humana y reconoce la información almacenada en la misma, para llevar a cabo un proceso, como se muestra en el diagrama a bloques de la figura 2.

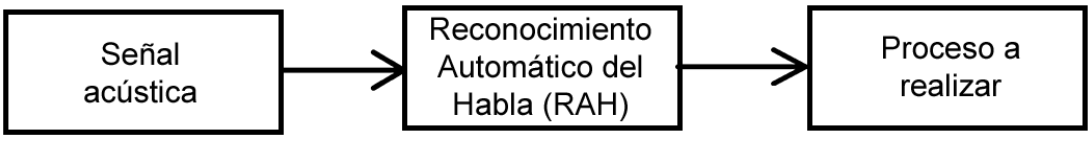

Fig. 2. Diagrama a bloques del ASR

La característica principal del PDS es la transformación de un tipo de señal de entrada en otro tipo de salida, esto se hace por medio de transductores, en éste caso con un micrófono, el cual transforma una señal acústica en una señal eléctrica, esta es transformada de analógica a digital por medio de un CAD para llevar acabo el tratamiento de la señal.

En los sistemas de comunicación de voz, la señal de voz se transmite, almacena y procesa de muchas maneras. Las técnicas ocupadas conducen a una amplia variedad de representaciones de la señal de voz. En general, hay dos cometidos principales en cualquier sistema:

1. Preservación del contenido del mensaje en la señal de voz.
2. Representación de la señal de voz en una forma que sea conveniente para la transmisión o el almacenamiento, o en una forma que sea flexible para que se puedan hacer modificaciones a la señal de voz sin degradar seriamente el contenido del mensaje.[1]

El presente trabajo se enfoca al segundo punto, el cual consiste en el tratamiento de la señal de voz en un sistema electrónico.

La investigación de este Proyecto surgió por el interés de desarrollar un robot para cirugía laparoscópica, controlado por un sistema de voz en tiempo real. Este es un proyecto en conjunto, el cual será probado en un simulador de cirugía laparoscópica (SIMULAP, fabricado en México) utilizando micro herramientas comerciales para su uso específico, como protocolos de entrenamiento laparoscópicos.

La presente tesis está dividida en 4 capítulos, a continuación, se describe brevemente el contenido de los mismos.

[1] R. Rabiner and R. W. Schafer "Digital Processing of Speech Signals". 1ª Edición, 1978. Pág. 2.

En el Capítulo 1 se define la metodología, el Planteamiento del problema, se establecen los objetivos generales y específicos, se realizan las preguntas de investigación, se formulan las hipótesis, se obtienen las variables de investigación y definición de las mismas. También en este capítulo se justifica el trabajo de la tesis y su viabilidad, por último se define los alcances y limitaciones del trabajo.

En el Capítulo 2 se desarrolla una Propuesta para poder realizar un algoritmo de RAH, esto por medio de una serie de técnicas y métodos matemáticos, con el objetivo de reconocer un número determinado de palabras aisladas, para poder realizar ciertas acciones ligadas a una entrada.

En el Capítulo 3 se realiza la implementación del algoritmo propuesto. Para esto se utiliza un sistema empotrado y distintos elementos.

En el Capítulo 4 se desarrollan las pruebas de funcionamiento, siguiendo una metodología que permite realizar experimentos y obtener los resultados esperados acorde al algoritmo, las conclusiones y recomendaciones.

Estado del Arte

En 1928 el físico e ingeniero sueco-estadounidense Harry Nyquist publicó el texto “Certain topics in telegraph transmission theory”, el cual fue comprobado en 1949 por el matemático, ingeniero eléctrico y criptógrafo estadounidense Claude E. Shannon en el trabajo denominado “Communication in the presence of noise”. La combinación de estos textos formula de manera conjunta lo que hoy conocemos como “Teorema de muestreo de Nyquist-Shannon”, este afirma que una señal analógica puede ser reconstruida, por muestras tomadas en intervalos de tiempos idénticos, “se debe muestrear la señal por lo menos dos veces en cada período o ciclo de su componente de frecuencia más alta.”[2]

Esto se expresa matemáticamente de la siguiente manera:

$$F_S > 2F_{max} = 2B$$

Donde F_S es la frecuencia de muestreo, F_{max} es la frecuencia máxima de la señal, y B es el ancho de banda, donde la mitad de su valor es frecuentemente llamada frecuencia de Nyquist.

En la figura 3 se muestra una señal analógica y su señal muestreada.

[2] B.P. Lathi. “Introducción a la teoría y sistemas de comunicación”. 1ª Edición, 2001. Pág. 96

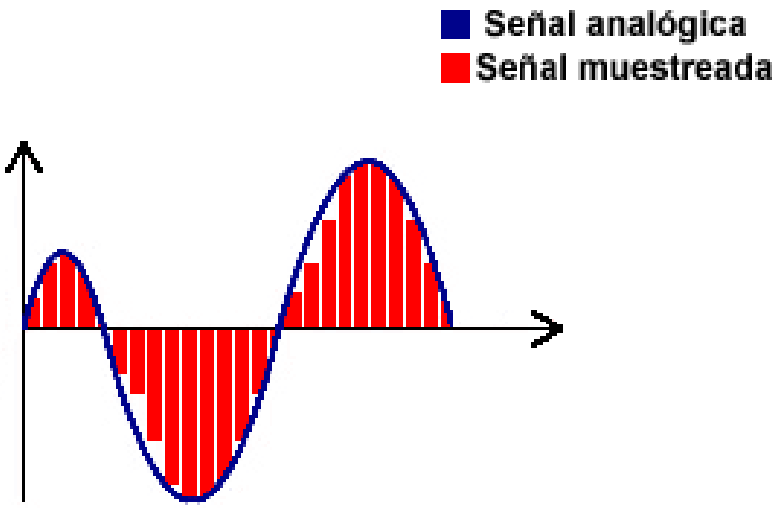

Fig. 3. Muestreo de una señal analógica

El reconocimiento Automático del Habla tiene probablemente sus orígenes con la invención del teléfono, por Antonio Meucci en 1854, patentado por Alexander Graham Bell en 1876, con el nombre que se le conoce hoy en día.

En 1877 Thomas Alva Edison inventó el fonógrafo, dispositivo que se usaba para grabar voz, por medio de cilindros de cera, con el fin de grabar discursos. Sin embargo, el uso de los mismos no se popularizo hasta finales de los años 1880, cuando fueron producidos en masa los cilindros de cera reutilizables.

El VODER (Voice Operating Demostrator; Demostrador operativo de voz) de Laboratorios Bell, fue el primer intento de sintetizar una voz humana electrónicamente, descomponiendo sus componentes acústicos. Fue inventado por Homer Dudley en 1937-1938.

El VODER sintetizó la voz humana, esto imitando los efectos del tracto vocal humano, en la figura 4, se muestra un diagrama a bloques del VODER.

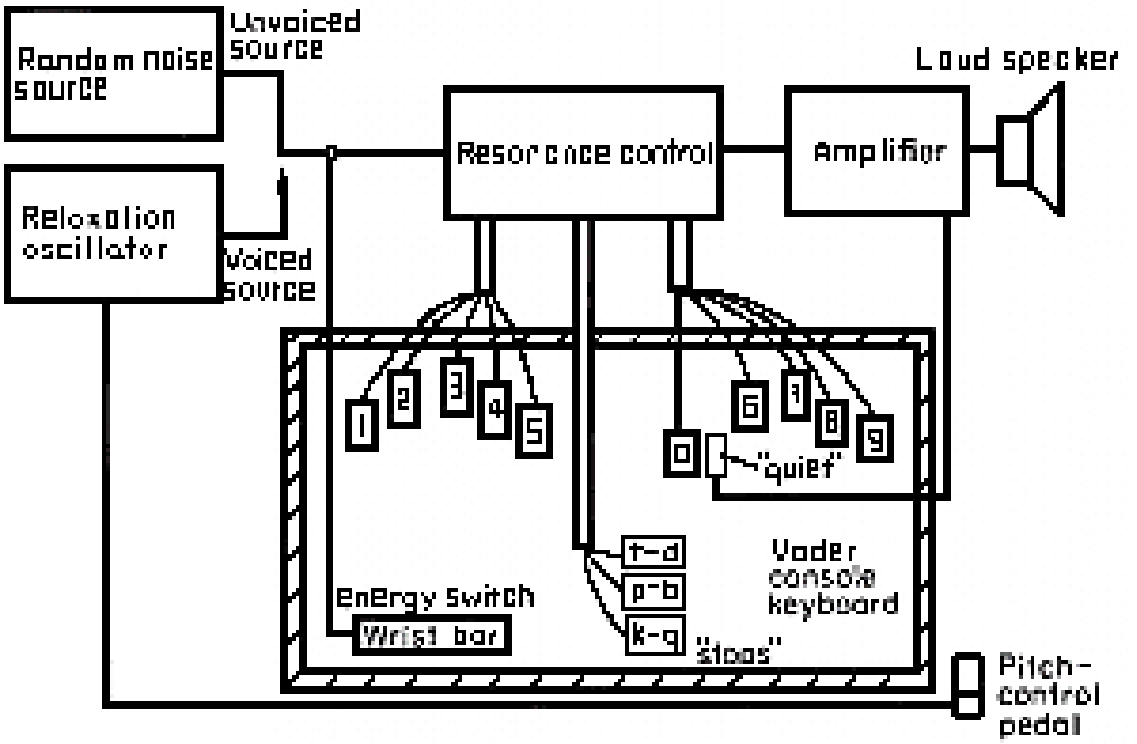

Fig. 4. Diagrama a bloques del VODER de Homer Dudley.

En la década de 1960 surge una explosión de trabajos relacionados principalmente con el reconocimiento de palabras aisladas.

En esta década IBM lanza “Shoebox”, el cual era un computador que podía realizar operaciones matemáticas y realizar reconocimiento de voz. Estaba limitado a reconocer 16 palabras habladas.

A finales de los 60´s la Agencia de Proyectos de Investigación Avanzados de Defensa (Advanced Research Projects Agency, ARPA; posteriormente en 1972 conocido como Defense Advanced Research Projects Agency, DAPRA), decide que la investigación del RAH había avanzado lo suficiente como para lanzar un reto, es decir, para ver si algún laboratorio de investigación podía construir un prototipo de un sistema funcional, que pudiera reconocer de forma precisa y confiable el habla de manera continua. Este proyecto comienza en 1971 con el nombre de ARPA-SUR (Advanced Research Projects Agency - Speech

Understanding System), este tuvo una duración limitada de 5 años, con metas como:

- Transformar fonemas (o silabas) en forma de palabras con la ayuda de la información lingüística.
- Usar estrés para identificar palabras de contenido semántico y límites de frase.
- Usar sintaxis para restringir secuencias de palabras.
- Usar semántica para restringir aún más las secuencias de palabras.

Aunque el proyecto ARPA-SUR no llegó a completar sus ambiciosos objetivos, las aportaciones del mismo contribuyeron de manera importante en la investigación de futuros proyectos.

Entre los años 80´s y 90´s surgen sistemas de amplio vocabulario y desde ese entonces los trabajos en este campo continúan siendo una rama de la investigación con objetivos y metas bien definidos.

Hoy en día se le considera el RAH como una rama interdisciplinaria de la computación, que desarrolla metodologías y tecnologías capaces de reconocer y traducir lenguaje hablado, a un lenguaje que pueda ser reconocido por un dispositivo electrónico, el RAH incorpora conocimientos e investigaciones de la computación, la lingüística y la electrónica.

A continuación, se muestra el diagrama del proceso del reconocimiento de voz.

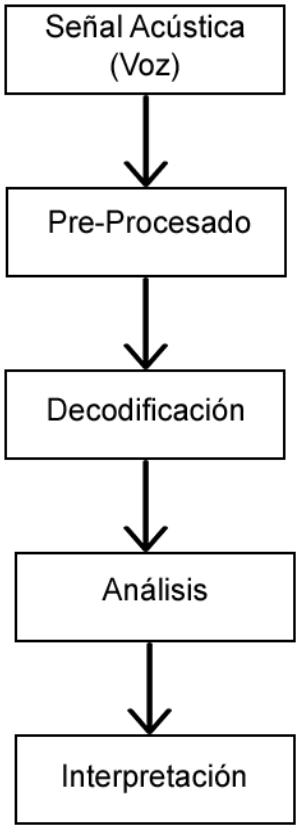

Fig. 5. Diagrama del proceso del reconocimiento de voz

La meta principal del Reconocimiento de la Voz es desarrollar técnicas y sistemas capaces de aceptar como entrada una señal hablada.[3]

En el trabajo “Mobile robot controlled by voice” presentado en el 2007, los autores Bojan Kuljic, Simon János y Szakáll Tibor, proponen la realización de un algoritmo eficiente de RAH para controlar un robot, con la meta final de desarrollar una plataforma de reconocimiento de voz, independiente del hablante y del idioma para uso con sistemas empotrados de bajos recursos. Utilizaron un método en el que dividieron los fonemas en distintos grupos según las características de los mismos como se muestra en la figura 6, el algoritmo utilizado en este trabajo

[3] Galindo Riaño, Pedro L. “Introducción al Reconocimiento de la Voz”. Primera Edición 1996. Pág. 1

funciona extrayendo fonemas individuales y formando palabras con los ya reconocidos.[4]

Group		Phoneme[a]
Vocals		*a e i o u*
Consonants	Plosives	*p t k b d g*
	Fricatives	*f h s š v z ž*
	Affricates	*c ć č đ dž*
	Nasals	*m n nj*
	Liquids	*l lj*
	Vibrant	*r*
	Semi-vowel	*j*

Fig. 6. Tabla de la división de fonemas en el algoritmo de "Mobile robot controlled by voice"

También en 2007 se presenta el trabajo titulado "Embedded speech recognition system for intelligent robot" en el cual los autores: Qingyang Hong, Caihong Zhang, Chen Yan y Xiao-Yang Chen, desarrollaron un sistema de RAH en un dispositivo empotrado basado en los Modelos Ocultos de Márkov (HHM; Hidden Markov Model), para la manipulación de un pequeño robot de juguete. Los autores realizaron su propio sistema empotrado basado en un chip de 16 bits, usaron un popular robot de juguete, el cual puede hacer 6 acciones que son controladas por motores eléctricos; basado en estas acciones se definieron los comandos para controlar los mismos, los cuales son "Gira a la derecha, "Gira a la izquierda" "Fuego" y "Baila" todos pronunciados en chino.

El diagrama a bloques del sistema de RAH se muestra en la figura 7.[5]

[4] Bojan K., Simon J., Szakáll T., "5th International Symposium on Intelligent Systems and Informatics". 2007. Págs. 189-192

[5] Hong, Q., Zhang, C. Chen X., Chen Y., Xia.Yang C., "14th International Conference on Mechatronics and Machine Vision in Practice". 2007. Págs. 35-38

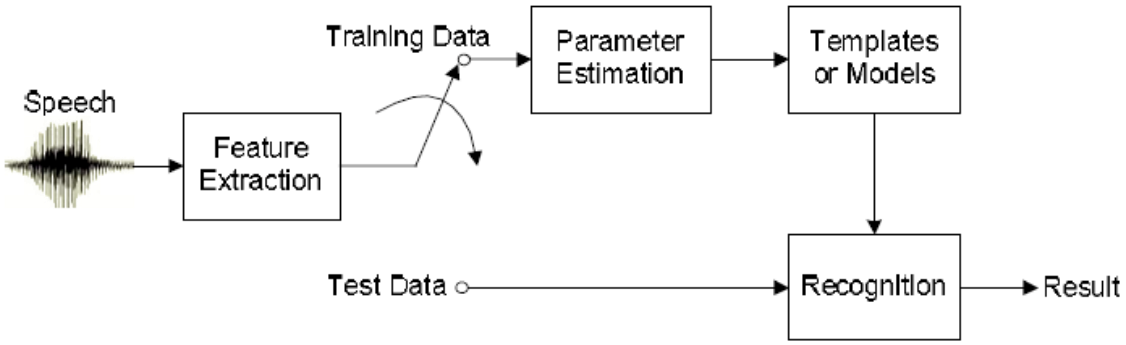

Fig. 7. Diagrama a bloques del sistema de RAH de "Embedded speech recognition system for intelligent robot"

Más tarde, en el año 2012 se presenta el trabajo titulado "The ASR Approach Based on Embedded System for Meal Service Robot" por los autores: Guo-Shing Huang, Sheng-Jr Yang, en el cual desarrollan un algoritmo de RAH para un robot de servicio de comida, esto para que un usuario pueda ordenar su comida más rápidamente, el algoritmo está basado en HMM y Coeficientes Cepstrales en la Frecuencia de Mel (Mel Frequency Cepstral Coefficients; MFCC).

En este trabajo se utilizó el formato MP3 para almacenar la voz y MATLAB para realizar las simulaciones, en la figura 8 se muestra el diagrama a bloques de la grabación y reproducción del sistema de RAH[6]

[6] Guo-Shing H., Sheng-Jr Y., "2012 International Symposium on Computer, Consumer and Control". 2012. Págs. 341-344

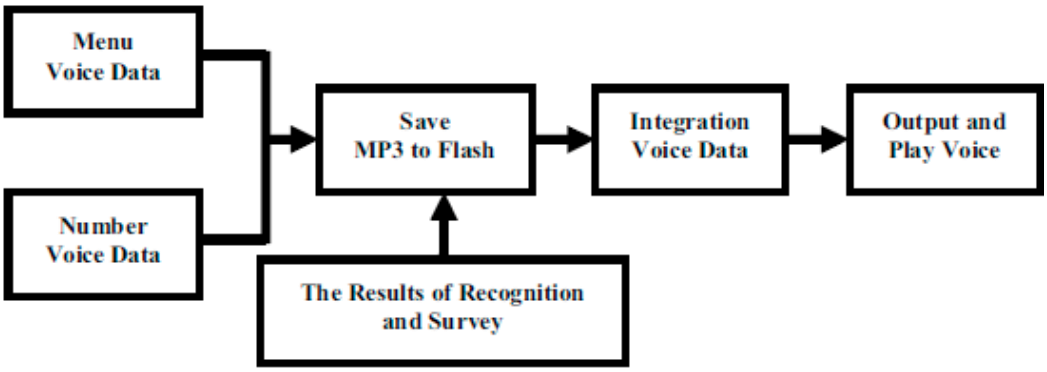

Fig. 8. Diagrama a bloques de la grabación y reproducción del sistema de RAH para "The ASR Approach Based on Embedded System for Meal Service Robot"

Tomando en cuenta los trabajos anteriores, en la presente tesis se propone un método de RAH, el cual es capaz de ser implementado en un sistema empotrado, para una aplicación específica orientada a la cirugía laparoscópica.

Capítulo 1 Metodología

La metodología es el conjunto de procedimientos y técnicas que son aplicados de manera ordenada para la realización de un estudio. Al respecto nos dice F. G. Ortiz Uribe, “La Metodología de la Investigación Científica es la disciplina que se ocupa de los principios y procedimientos, técnicas y procedimientos, técnicas e instrumentos de conocimientos para descubrir la verdad y enseñarla.”[7]

En el presente trabajo la metodología seguida, es el diseño y la implementación de un algoritmo de RAH, para esto es necesario conocer el problema para determinar los pasos a seguir. Se realizó la investigación correspondiente, que consta del planteamiento del problema, los objetivos, las preguntas de investigación, la hipótesis, los alcances y limitaciones, la viabilidad de la investigación y la justificación de la misma.

1.1 Planteamiento del Problema

La cirugía laparoscópica que se practica actualmente en nuestro país, es una mejora significativa a comparación de la cirugía tradicional “abierta”, en tiempo de recuperación, costos y estética. Sin embargo, aun cuando la cirugía laparoscópica es mejor por lo antes expuesto, se requiere de un cirujano en jefe (el que realiza la cirugía), un ayudante médico que maneja la cámara laparoscópica, el anestesiólogo y las enfermeras asistentes. Lo que resulta algo problemático al haber varias personas en la sala de operaciones.

[7] Ortiz Uribe. “Diccionario de Metodología de la Investigación Científica”. Segunda Edición, 2008. Pág. 142.

En el presente trabajo se propone realizar un brazo robótico controlado por comandos de voz para cirugía laparoscópica, con la finalidad de brindar a la medicina y en particular a los cirujanos especialistas, una herramienta que disminuya los errores humanos, mejore la precisión y reducción de los costos de la operación y mejore el tiempo de recuperación de los pacientes.

Mediante un sistema de control por voz, el cirujano en jefe podría manipular un Brazo Robótico para la Cirugía Laparoscópica (BRCL) por medio de comandos, lo cual optimizará la interacción humano-computadora, dando como resultado una mayor precisión durante el proceso quirúrgico.

1.2 Objetivos Generales y Específicos

General

- Diseñar e implementar un sistema de control por voz para manipular un BRCL.

Específicos

- Identificar y evaluar artículos científicos donde se reporte el desarrollo e implementación de algoritmos del RAH
- Diseñar el algoritmo de RAH para el sistema empotrado.
- Determinar el número de comandos que reconocerá el sistema de RAH.
- Manipular mediante comandos de voz un BCRL.
- Evaluar el desempeño del sistema de RAH
- Reportar resultados y avances obtenidos.

1.3 Preguntas de investigación

- ¿Qué se necesita para desarrollar el sistema de control para reconocimiento de voz?
- ¿Cuáles son las etapas para desarrollar el algoritmo de control?
- ¿Cómo verificar la funcionalidad del algoritmo?

1.4 Hipótesis

1. Es posible manipular un brazo robótico para cirugía laparoscópica por medio de un sistema de control utilizando comandos de voz.
2. El algoritmo diseñado para el sistema de control por voz puede ser implementado en cualquier sistema empotrado.

1.5 Variables

De la Hipótesis 1:

- Manipulación
- Brazo robótico.
- Cirugía laparoscópica.
- Sistema de control.
- Comandos de voz.

De la Hipótesis 2:

- Algoritmo diseñado.
- Sistema empotrado.

1.6 Definición de Variables

Manipulación

Operar uno o varios objetos de manera manual o a través de herramientas quirúrgicas.

Brazo robótico

Es un brazo mecánico, normalmente programable de n grados de libertad, con funciones parecidas a las de un brazo humano, este puede ser la suma total del mecanismo o puede ser parte de un robot más complejo. Las partes de estos manipuladores o brazos son interconectadas a través de articulaciones que permiten, tanto un movimiento rotacional (tales como los de un robot articulado), como un movimiento traslacional o desplazamiento lineal.

Cirugía laparoscópica

La laparoscopía es una técnica quirúrgica que permite observar el interior del abdomen para establecer un diagnóstico y también se emplea para realizar una operación. Para ello se realizan pequeñas incisuras en la pared abdominal y, a través de ellas se introducen cámaras, pinzas, cuchillas y otros aparatos muy pequeños que permiten manipular las vísceras internas sin tener que abrir el abdomen del todo.

Sistema de control

Un sistema de control es una interconexión de componentes que forman una configuración del sistema que proporcionará una respuesta deseada.

Comandos de voz

Órdenes o instrucciones dadas por uno o varios usuarios a un sistema de control desde el habla humana a través de transductores.

Algoritmo

Es un conjunto pre-escrito de instrucciones o reglas bien definidas, ordenadas y finitas que permiten llevar a cabo una actividad mediante pasos sucesivos que no generen dudas a quien deba hacer dicha actividad.

Sistema empotrado

Un sistema empotrado o embebido, es un sistema electrónico del tipo SoC (System on a Chip) ya que su característica principal es que todos los componentes electrónicos están integrados en una sola placa, son sistemas diseñados para propósitos específicos, poseen ALU, memoria, dispositivos de entrada/salida y pueden o no tener un sistema operativo, además de ser programables.

1.7 Justificación

La motivación de realizar un brazo robótico controlado por 7 comandos de voz para cirugía laparoscópica es, brindar a la medicina y en particular a los cirujanos especialistas, una herramienta que facilite su trabajo, con un menor riesgo, mayor precisión y mejor manipulación. Mediante este trabajo de tesis se contribuirá al desarrollo tecnológico en el campo médico de México, con tecnología nacional.

El impacto de esta tesis será positivo para la técnica de la cirugía laparoscópica, que ha mostrado reducir las complicaciones durante la cirugía, tiempo de recuperación y estética del paciente, en comparación con la cirugía abierta; en un futuro cercano los costos de este tipo de operación irán disminuyendo.

Debido a los avances tecnológicos y poco desarrollo en México, se depende de la tecnología extranjera. El poder diseñar y construir herramientas en México, para el campo de la cirugía laparoscópica brinda una solución a las dificultades que se presentan en el quirófano.

1.8 Viabilidad de la investigación

Esta investigación se puede realizar y llevar a la práctica ya que dentro de las instalaciones del Laboratorio de Sistemas Robóticos (SIRO) de la FCC-BUAP, se cuenta con todos los elementos necesarios para realizar la misma, tales como: un sistema empotrado con las características necesarias para realizar el procesamiento digital de señales, un micrófono, una tarjeta de audio USB, una memoria SD, un teclado, un mouse, un monitor, jumpers, protoboard, push button y leds.

Para llegar a cumplir los objetivos de este trabajo es necesario consultar diversas fuertes de información, como libros, artículos científicos y bases de datos fidedignos.

1.9 Alcances y limitaciones

Se implementó un algoritmo de RAH en el lenguaje de programación Python, que permitió digitalizar la señal de la voz humana, procesarla y obtener

una respuesta en el sistema empotrado utilizado. Lo cual permitió capturar señales de voz de entrada, guardarlas en una base de datos y posteriormente compararlas con las nuevas señales de entrada.

El sistema de reconocimiento de voz es capaz de manipular un prototipo de BRCL.

Como limitaciones el sistema empotrado utilizado tiene una capacidad reducida de procesamiento, no cuenta con CAD ni con CDA, por lo cual se tiene que adaptar una tarjeta de audio USB. No cuenta con un dispositivo de almacenamiento de datos, lo que obliga a utilizar una memoria SD para el almacenamiento de los mismos. Mientras el sistema esté alimentado con la batería el tiempo de funcionamiento dependerá de los procesos realizados en el mismo.

El sistema empotrado cuenta con un número determinado de pines de entrada y salida, lo cual limita las señales de control que se utilizan.

Capítulo 2 Diseño del algoritmo

El sistema de control por voz, es un sistema de control realimentado o sistema de control en lazo cerrado.

En un sistema de control en lazo cerrado, se alimenta al controlador la señal de error de actuación, que es la diferencia entre la señal de entrada y la señal de realimentación (que puede ser la propia señal de salida o una función de la señal de salida y sus derivadas y/o integrales), con el fin de reducir el error y llevar la salida del sistema a un valor deseado. El término control en lazo cerrado siempre implica el uso de una acción de control realimentado para reducir el error del sistema.[8]

Una señal se define como cualquier magnitud física que varía con el tiempo, el espacio o cualquier otra variable o variables independientes. Matemáticamente, describimos una señal como una función de una o más variables independientes.[9]

El proceso de convertir una señal analógica en una señal digital, es llamado Conversión digital (CD; Digital Convertion, A/D), y el dispositivo encargado de hacerlo es llamado convertidor analógico/digital (CAD; Analog-to-Digital Converter, ADC)[10]

[8] K. Ogata. "Ingenieria de control moderna". 5ª Edición, 2010. Pág. 7.
[9] J. G Proakis. D.G. Manolakis, "Tratamiento digital de señales". 4ª Edición, 2007. Pág. 2.
[10] H. Huang., Acero, A., H. W. Hon, "Spoken Language Processing - A Guide to Theory, Algorithm, and System Development".1ª Edición, 2001. Pág. 245.

En la figura 9 se muestra un sistema básico de control en lazo cerrado.

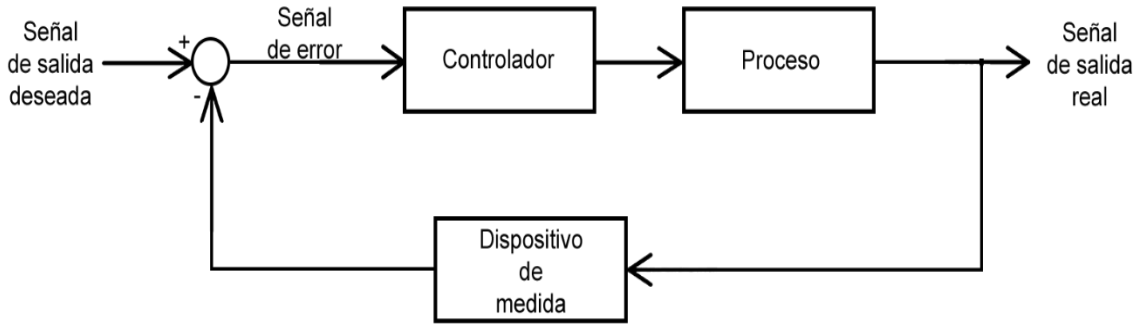

Fig. 9. Sistema de control en lazo cerrado

El sistema de control en lazo cerrado, se ha desarrollado implementando un algoritmo de RAH, para esto se utilizó el sistema empotrado Raspberry pi 3, al cual se le adaptó una memoria SD con un sistema operativo precargado, en este caso se utilizó Ubuntu MATE de 64 bits en su versión 16.04, una tarjeta de sonido USB Manhattan (que funciona como CAD y CDA) y un micrófono para la captura de audio. En la figura 10 se muestra el sistema empotrado y sus componentes.

Fig. 10. Sistema Empotrado para la obtención de señales de audio.

La Raspberry Pi es una computadora del tamaño de una tarjeta de crédito diseñada y fabricada en Reino Unido con la intención inicial de proporcionar un dispositivo informático barato para la educación. Sin embargo, desde su lanzamiento ha crecido mucho más allá de la esfera de la academia.[11]

Existe una gran variedad de algoritmos de RAH para diferentes tipos de aplicaciones, en el presente trabajo se tomaron algunos como referencia para hacer una propuesta de diseño propio.

Como primera referencia se tomó el algoritmo de RAH de la figura 11. En esta comunicación se presenta un sistema empotrado sobre FPGA de reconocimiento de voz que aplica el algoritmo LPC (Linear Predictive Coding).[12]

[11] A. K. Dennis. "Raspberry Pi Computer Architecture Essentials". 1ª Edición. Pág. 1.
[12] Balosa, Jesús., Crespo, Francisco J., Barriga, Angel. "Sistema empotrado de reconocimiento de voz sobre FPGA". Pág. 1.

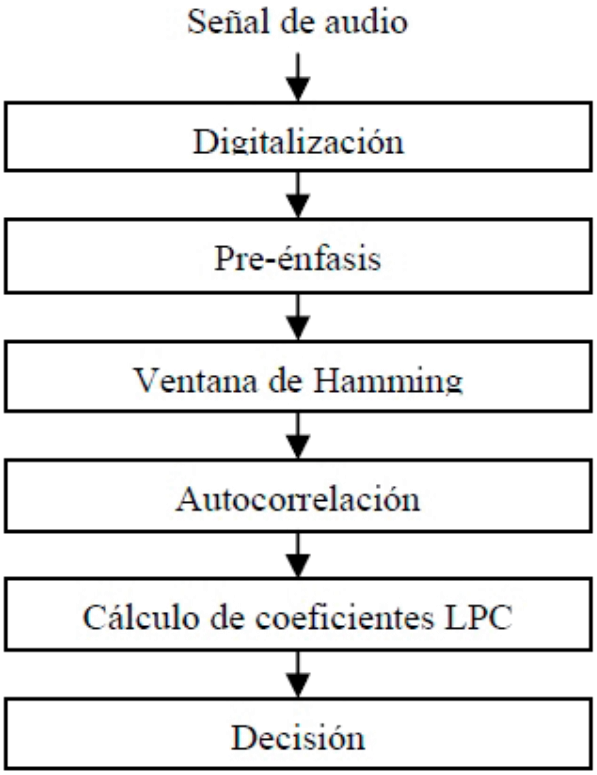

Fig. 11. Algoritmo LPC (Linear Predictive Coding), de la primera referencia

Como segunda referencia se tomó el algoritmo de RAH mostrado en la figura 12.

Se creó una Interfaz Gráfica de Usuario (GUI, Graphical User Interfaces) en MATLAB, para el reconocimiento de palabras aisladas (comandos), utilizando un micrófono multimedia y la tarjeta de audio de una computadora personal. Para la caracterización de las palabras, se aplicaron técnicas como: predicción lineal, coeficientes Cepstrum y polinomios ortogonales.[13]

[13] Villarreal Robles, G., Olivera Reyna, R. "Reconocimiento de comandos de voz utilizando técnicas de PDS aplicadas a robótica". 2010. Pág. 1.

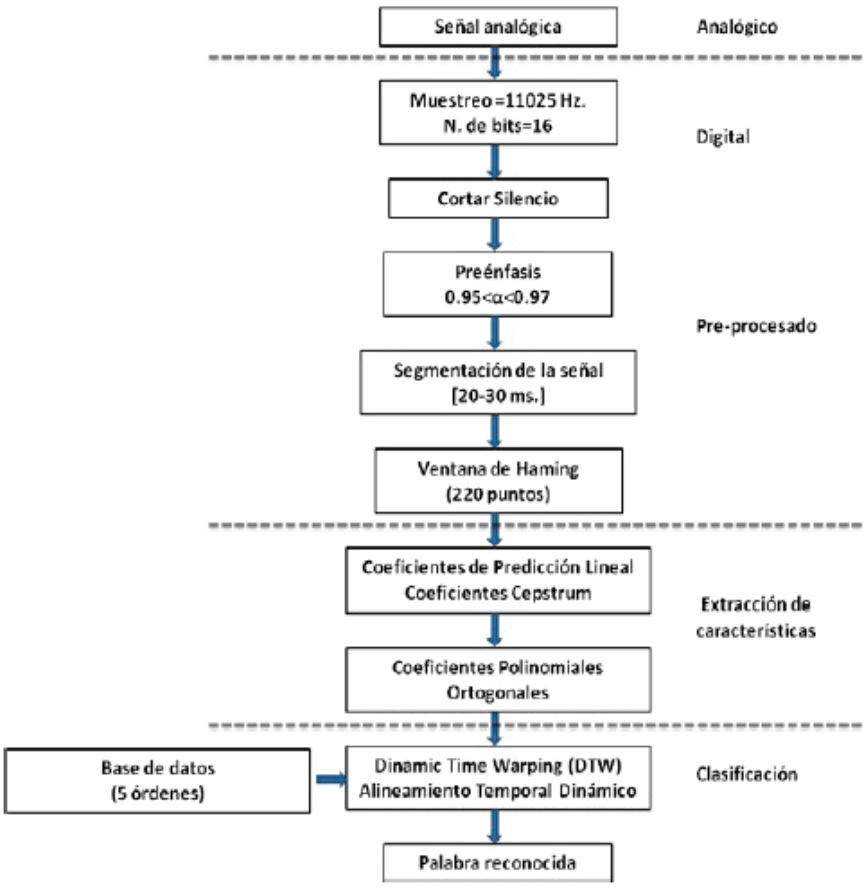

Fig. 12. Diagrama a bloques del sistema reconocedor de palabras, de la segunda referencia

En base a estas propuestas se ha diseño un algoritmo de RAH propio que cumpla con los requerimientos de este proyecto.

El algoritmo de RAH consta de 9 etapas para la captura y procesamiento de señales.

En la figura 13 se muestran las etapas que se propusieron para la obtención y tratamiento de la señal de audio, estas se visualizarán de manera práctica en el Capítulo 3 el cual corresponde al desarrollo y pruebas de funcionamiento.

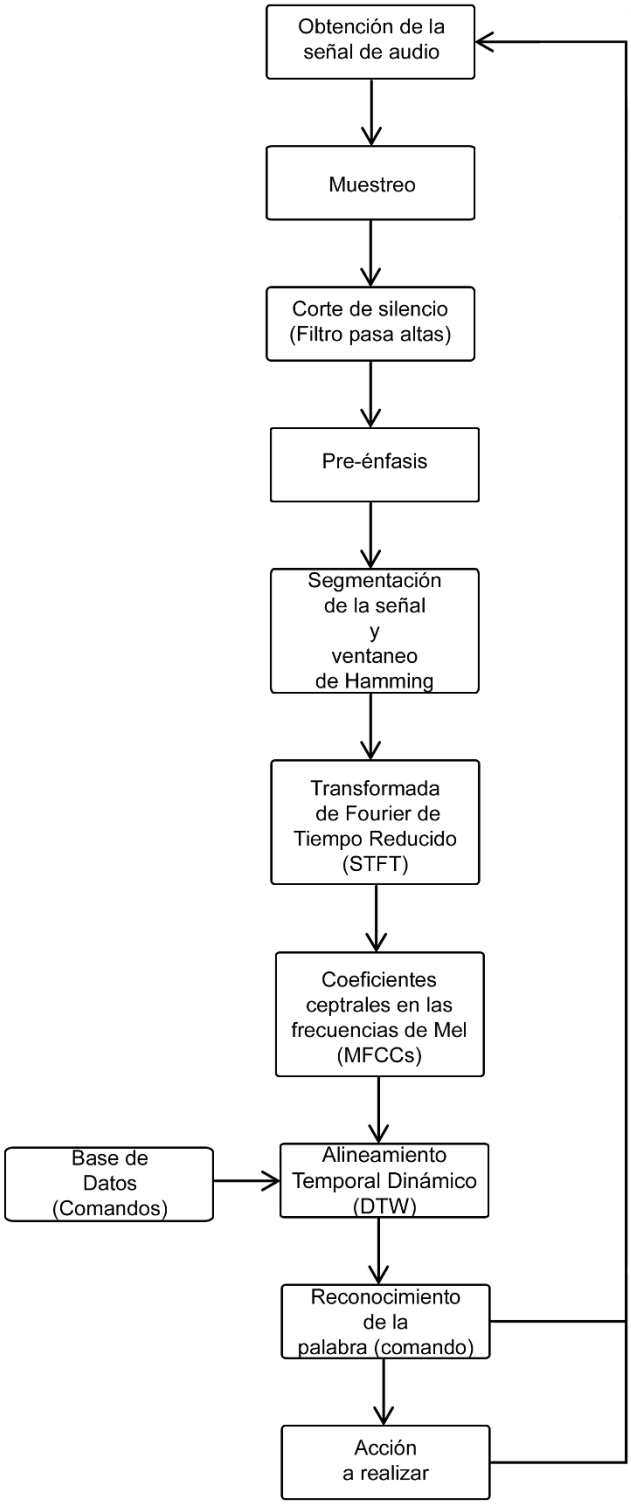

Fig. 13. Algoritmo para la adquisición de datos y tratamiento de la señal de audio (Autoría propia).

A continuación, se muestra la información teórica que sustenta cada una de las etapas del algoritmo propuesto para la obtención y procesamiento de la señal de audio.

2.1 Obtención de la señal de audio y muestreo

El primer paso es obtener la señal analógica, para lo cual un usuario enunciará una palabra a través de un micrófono, con lo cual el sistema podrá entender la misma señal para poder hacer el resto del procesamiento de manera digital.

En la figura 14, se muestran las etapas necesarias para la obtención de la señal de audio.

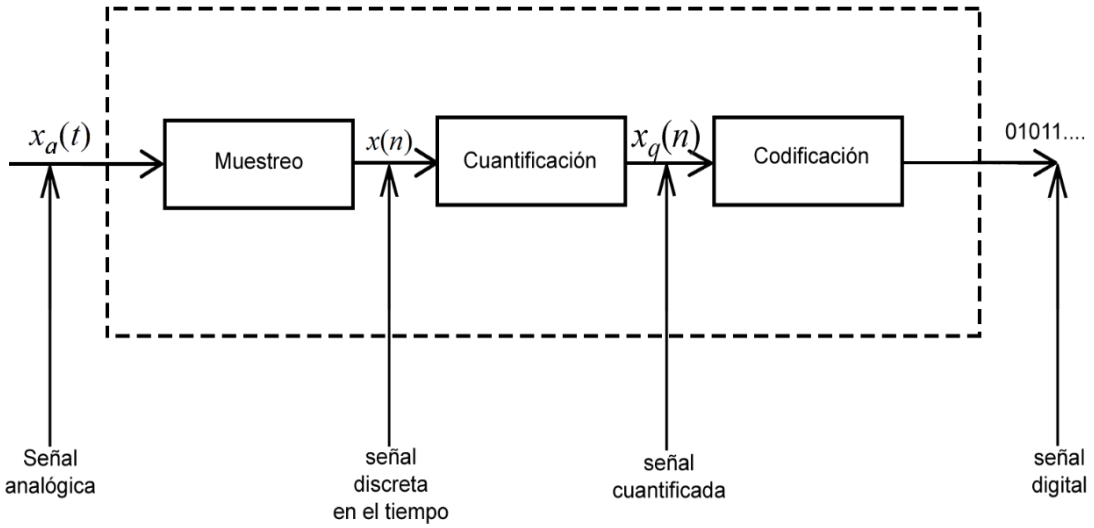

Fig. 14. Diagrama de bloques del CAD

- *Muestreo:* Este paso consiste en la conversión de una señal continua en el tiempo en una señal discreta en el tiempo obtenida mediante la toma de

"muestras" de la señal continua en el tiempo en instantes discretos de tiempo. Por tanto, si $x_a(t)$ es la entrada del muestreador, la salida será $x_a(nt) \equiv x(n)$, donde T es el intervalo de muestreo.

- *Cuantificación:* En este paso se realiza la conversión de una señal de valores continuos tomados en instantes discretos de tiempo en una señal de valores discretos en instantes de tiempo discretos (es decir, es una señal digital). El valor de cada muestra de la señal se representa mediante un valor seleccionado dentro de un conjunto finito de posibles valores. La diferencia entre la muestra no cuantificada $x(n)$ y la salida cuantificada $x_q(n)$ es el error de cuantificación.

- *Codificación:* En el proceso de codificación, cada valor discreto $x_q(n)$ se representa mediante una secuencia binaria de b-bits.[14]

Para convertir la señal de audio analógica a digital, se requirió de un micrófono, el cual actúa como transductor, transformando la voz en señales eléctricas, esto con el fin de obtener la señal analógica de entrada y convertirla en una señal digital, lo cual se hace transformando las señales en el tiempo continuo a discreto.

El muestro más común utilizado es el periódico, que consiste en tomar muestras de la misma señal en intervalos iguales.

$$x(n) = x_a(nt)$$

Donde $x(n)$ es la señal obtenida de las muestras hechas en la señal analógica, y $x_a(t)$es dicha señal en cada t segundos. En la imagen 15 se muestra una señal analógica y su señal muestreada.

[14] J. G. Proakis., D.G. Manolakis. Op. Cit. Pág. 17.

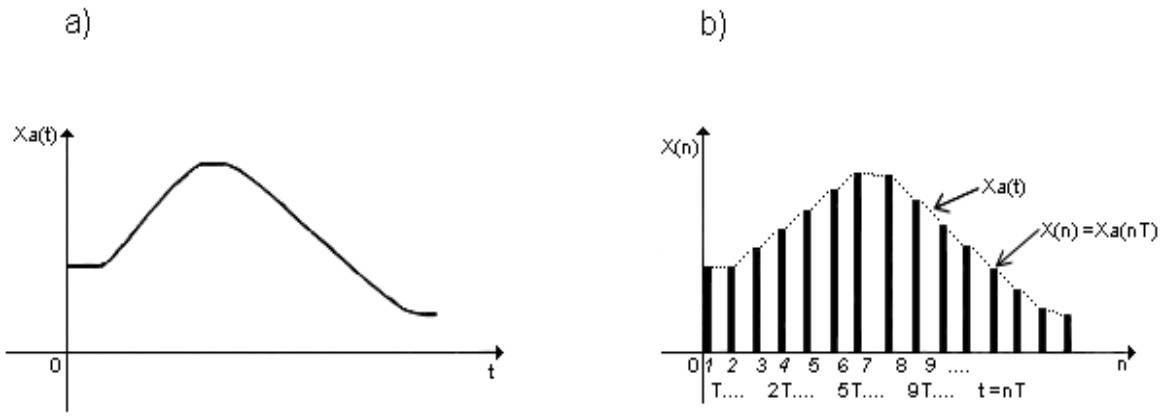

Fig. 15. Muestreo de señal analógica:
a) Señal Original, b) muestreo de la señal

La frecuencia de la voz humana se mide en Hercios (Hz) y va de los 250Hz a 3.5KHz, habiendo casos espécielas que alcanzan hasta los 5KHz.

Para obtener la frecuencia de muestreo correcta, se tomó en cuenta el teorema de Nyquist, el cual nos dice: para que una muestra analógica pueda ser reconstruida sin pérdida de información de manera digital, se toma la máxima frecuencia de la señal analógica y se muestrea al menos al doble de la misma.

$$F_{max} = B\ en\ Xa(T)$$
$$\therefore F_s \geq 2F_{max} \equiv 2B$$

Donde F_{max} es la frecuencia máxima de la señal analógica, y F_s es la frecuencia de muestreo, y B el ancho de banda, de esta manera si sustituimos los valores de nuestra ecuación obtendremos nuestra frecuencia de muestreo.

Ejemplo: tomando como la frecuencia máxima de la voz que es de 5KHz.

$$F_s \geq 2(5Khz) \equiv 10KHz$$

Esta es la frecuencia mínima de muestro para evitar el *Aliasing*, el cual es el efecto en el cual las señales continuas se tornen indistinguibles, cuando se muestrean digitalmente.

Si la tasa de muestreo no satisface el criterio de Nyquist, los períodos adyacentes del espectro analógico se superpondrán, lo que provocará un espectro distorsionado. Este efecto, llamado distorsión de Aliasing, es bastante serio porque no se puede corregir fácilmente una vez que ha ocurrido.[15]

En la figura 16, se muestra una señal analógica la cual está siendo muestreada de manera incorrecta lo que provoca el efecto de Aliasing.

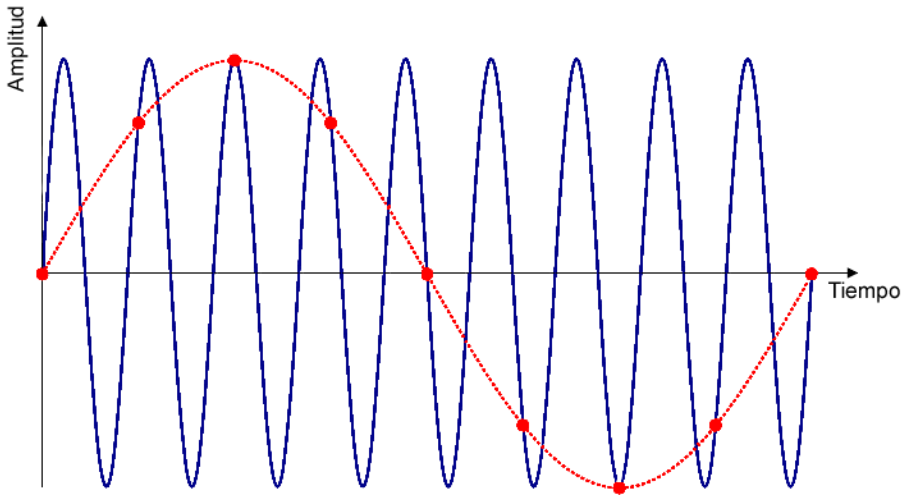

Fig. 16. Efecto de Aliasing

Para evitar el efecto anteriormente explicado, la frecuencia de muestreo usada en este caso será de 11025Hz, esto ya que es más del doble de lo requerido y lo cual nos dará una reconstrucción digital sin perdidas de información, además el formato utilizado, WMA (Windows Media Audio) maneja la frecuencia propuesta como uno de sus estándares.

[15] V.K. Madisetti, D. B. Williams. "Digital Signal Processing Handbook". 3ª Edición, 1999. Pág. 39.

Una vez terminado el muestreo, el siguiente paso es la cuantificación, es decir, para cada una de nuestras muestras existe un número de posibles valores digitales, este valor está dado por

$$2^n - 1$$

Donde n es el número de bits, y se le resta un 1 ya que los bits empiezan desde 0.

El número de bits nos da la resolución con la que se está trabajando, para efecto práctico se tomó una resolución de 16 bits, que es la estándar usada en formatos de audio digital.

Dados los 16 bits, la ecuación quedaría como:

$$2^{16} - 1 = 65{,}535$$

Así que tendremos 65,535 posibles intervalos por muestreo, dicho de otro modo, cada segundo de grabación tendremos 11025 muestras cada una de ellas con 65,535 posibles intervalos, lo que nos dará como resultado una grabación de audio de calidad.

El último paso de esta etapa es la codificación del audio, donde la señal analógica es transformada a digital, para que pueda ser comprendida por un sistema electrónico.

Se utilizó un solo canal, sonido monoaural o monofónico (Comúnmente abreviado como “mono”) para capturar las señales de audio, es decir, en este caso se usó un solo micrófono para capturar dichas señales.

Con esto se determinaron los parámetros que se usaron en nuestro algoritmo:

Canales: 1(mono)
Frecuencia de muestreo: 11025Hz
Resolución: 16 bits

Con estos valores es posible codificar la señal analógica a una señal digital y concluir la primera etapa de nuestro algoritmo de RAH.

2.2 Corte de silencio

Posterior a la obtención de la señal de audio se necesitó aplicar un filtro digital para cortar el silencio, ya que al capturar audio existe un pequeño lapso de silencio o de frecuencias bajas (por debajo de los 250 Hz), las cuales no son necesarias procesar, se utilizó lo que comúnmente se denomina filtro pasa altas.

Un filtro pasa altas se diseña para dejar pasar las frecuencias superiores a su frecuencia de corte F_c[16]

En la figura 17, se muestra una gráfica del filtro pasa altas real, donde el eje x son las frecuencias medidas en Hz (F), y el eje y es la ganancia que es la razón del voltaje de salida (Vout) contra el voltaje de entrada (Vin).

[16] H. Huang., Acero., A., H.W. Hon. Op. Cit. Pág. 639.

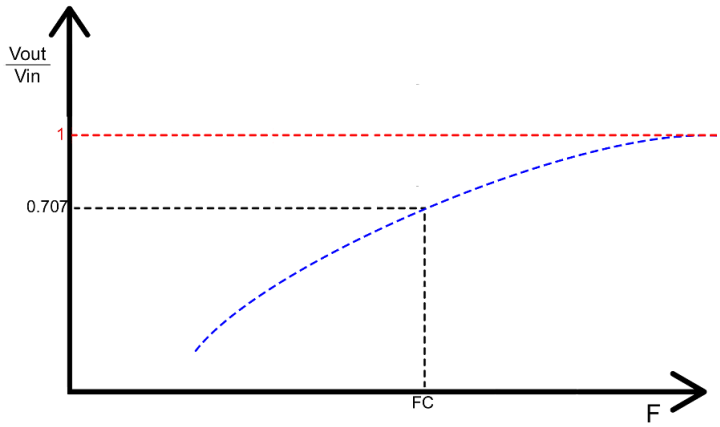

Fig. 17. Corte de silencio (filtro pasa alta)

2.3 Preénfasis

El preénfasis es un proceso diseñado para incrementar magnitudes de frecuencias que son usualmente altas con respecto a magnitudes usualmente bajas, ya que el tracto vocal no filtra de manera efectiva las frecuencias altas.

La señal de voz digitalizada, se pasa a través de un sistema digital de bajo orden (en realidad un filtro FIR de primer orden) para aplanar espectralmente la señal y hacerla menos susceptible a los efectos de precisión finita que se darán posteriormente en el procesamiento de la señal.[17]

Un filtro FIR (Finite Impulse Response) es un tipo de filtro digital, a menudo se lleva a cabo mediante convolución, en lugar de recurrencia. Su respuesta de impulso (La reacción de cualquier sistema dinámico en respuesta a algún cambio externo) es de duración finita, porque el tiempo finito se establece a cero.

El filtro de preénfasis está definido por la siguiente función:

[17]Milan G. Mehta. “Speech Recognition System”. 1996. Pág. 25.

$$H(z) = 1 - 90z^{-1} \quad \textbf{(3)}$$

2.4 Segmentación de señal y ventana de Hamming

En el RAH es común dividir la señal en segmentos para lograr una señal estacionara, es decir una señal que es constante en sus parámetros estadísticos sobre el tiempo. Esto ya que normalmente las señales del habla son no estacionarias.

La señal debió ser segmentada en intervalos de 20 a 30ms donde se considera que es una señal estacionaria cuasi periódica. Esta segmentación es aplicada ya que si utilizamos una más corta no tenemos suficientes muestras para obtener una estimación espectral adecuada, y si es más larga la señal cambia demasiado a lo largo del segmento.

En la figura 18 se muestra una señal arbitraria y como es dividida en segmentos. Cada segmento comparte la primera parte con el segmento anterior y la última parte con el segmento siguiente.

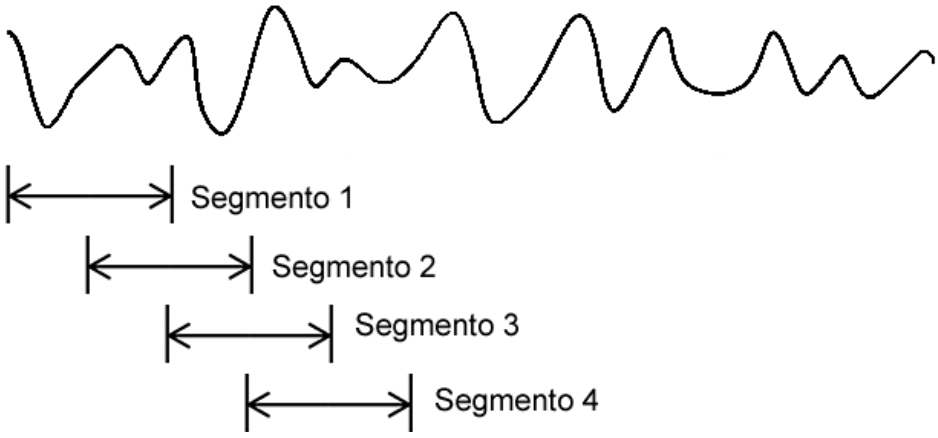

Fig. 18. Señal arbitraria dividida en segmentos

Una vez que la señal ha sido segmentada, es posible aplicarle una función a cada uno de los segmentos, en este caso se aplicó la función ventana de Hamming o de coseno elevado, esto con el fin de minimizar las oscilaciones causadas por el fenómeno de Gibbs.

Cuando se usa una serie de Fourier para aproximar una función con una discontinuidad de salto, se produce un sobre impulso en los puntos de discontinuidad. Este fenómeno fue notado por Michelson y explicado por Gibbs en 1899. Este fenómeno se conoce como el efecto de Gibbs.[18]

Una función ventana o simplemente llamada ventana, es solo un vector de longitud M, multiplicado por los segmentos originales, esto para alterar su amplitud de manera selectiva.

Las funciones de ventana son señales que están concentradas en el tiempo, a menudo de una duración limitada. Mientras algunas funciones como la triangular, Kaiser, Barllet y la esferoidal prolata aparecen ocasionalmente en sistemas de procesamiento digital de voz, las funciones como la rectangular, Hanning y Hamming son las más utilizadas en estos sistemas.[19]

Los lóbulos laterales en el dominio de la frecuencia son una manifestación de las discontinuidades en el dominio del tiempo en los bordes de una ventana rectangular y se pueden aliviar mediante el uso de ventanas que no contienen discontinuidades agudas y se deslizan suavemente hacia cero, como el coseno ventanas de Hamming.[20]

---

[18] Gu xiaohong, Cai jinhui. "Proceedings of the 7th World Congress on Intelligent Control and Automation". 2008. Págs. 7564 – 7566.

[19] H. Huang., A. Acero., H.W. Hon, Op. Cit. Pág 230

[20] S. V. Vaseghi, "Multimedia Signal Processing: Theory and Applications in Speech, Music and Communications". 1ª Edición, 2007. Pág. 129.

La ventana de Hamming tiene una forma sinusoidal, está dado por la función de coseno elevado y como resultado da un pico alto, pero lóbulos laterales bajos, como se muestra en la figura 19. Esta ventana no llega a 0 en sus lóbulos, por lo que al aplicarla aún existen ligeras discontinuidades, sin embargo, hace un mejor trabajo en relación a otras ventanas al cancelar el lóbulo lateral más cercano.

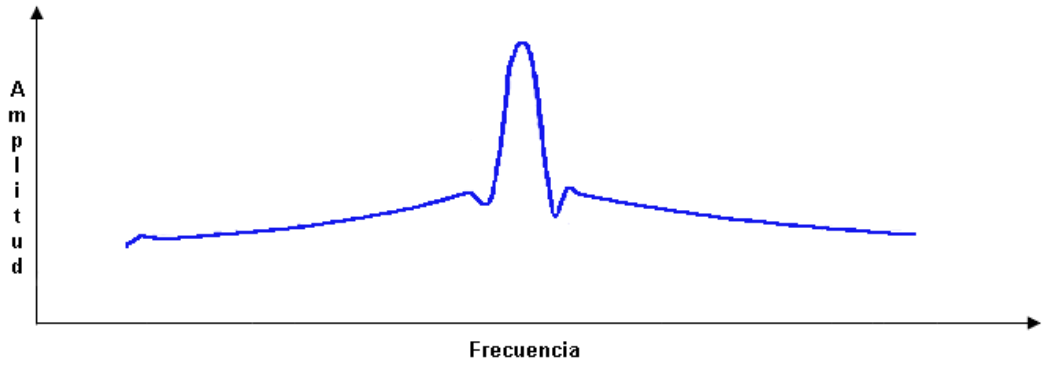

Fig. 19. Ventana de Hamming

La venta de Hamming está definida por:

$$v(n) = \sum_{i=0}^{N-1} [a_0 - a_1 \cos \frac{2\pi n}{N-1} \quad \textbf{(4)}$$

Donde $a_0 = 0.53836$ y $a_1 = 0.46164$

2.5 Transformada de Fourier de Tiempo Reducido

Una vez que la señal ha sido segmentada y enventada, se le aplica una Transformada de Fourier a cada segmento de la misma, para calcular el espectrograma, es decir, una representación bidimensional que muestra el tiempo en su eje horizontal y la frecuencia en su eje vertical, este proceso también es

conocido como Transformada de Fourier de Tiempo Reducido (Short-time Fourier transform, STFT)

En aplicaciones de la vida real y de tiempo real, las señales tienen duración finita y antes de la aplicación de la transformada de Fourier estas comúnmente son divididas en segmentos de longitud relativamente cortos, por dos principales razones:

- La existencia de límites en la demora tolerable del tiempo en sistemas de comunicación, y el requerimiento de reducir la complejidad computacional tanto como sea posible, esto implica que la señales necesitan ser divididas en segmentos de longitud relativamente cortos.
- La teoría de Fourier asume que las señales son estacionarias; esto significa que las estadísticas de la señal, tales como la media, la potencia, y la potencia del espectro son invariantes en el tiempo. La mayoría de las señales de la vida real, tales como el habla, la música, la imagen, y el ruido son no estacionarias en su amplitud, potencia, composición espectral y otras características que cambian continuamente en el tiempo.[21]

El espectrograma es una gráfica de la variación del espectro de la magnitud de tiempo reducido (o poder) de una señal de tiempo. La señal es dividida dentro de segmentos ventaneados, sobrepuestos de una apropiada duración reducida (aproximadamente 25ms para señales de audio), cada segmento es transformado con la FFT, y los vectores de frecuencia de la magnitud se amplían y trazan con la representación del eje vertical representando la frecuencia y el eje horizontal el tiempo. Los valores de la magnitud están codificados por colores, el color negro representa el valor más bajo, mientras que el blanco el valor más grande.[22]

[21] S. V. Vaseghi. Op. Cit. Págs. 57-58.
[22] Op. Cit. Pág. 68.

En la figura 20, podemos observar un espectrograma de una señal de voz, donde se aprecia la intensidad según los colores blancos y negros.

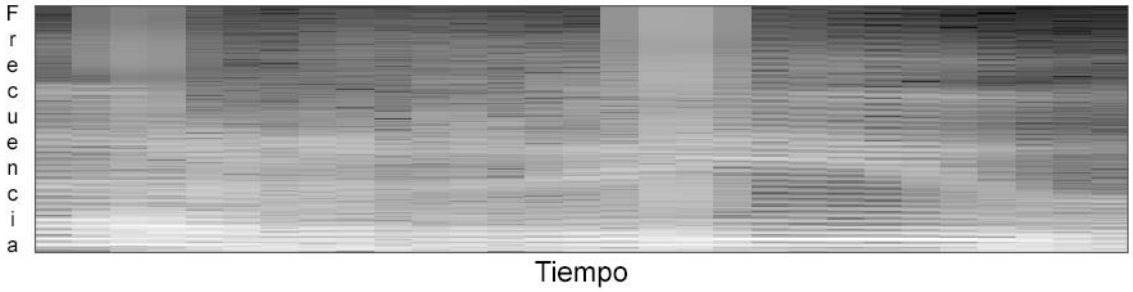

Fig. 20. Espectrograma de una señal de voz

2.6 Coeficientes Cepstrales en la Frecuencia de Mel

El método MFCC, es uno de los métodos más utilizados para extraer características de voz.

Antes de usar este método tenemos que calcular los filter Banks o bancos de filtros, para nuestro caso se denominan Mel scale filter bank spacing (espaciado entre bancos de filtro de escala Mel). En este paso se multiplica la señal por un filtro de bancos triangulares, estos triángulos están en la escala de Mel. El ancho de banda se determina por la distribución de la frecuencia del centro de cada filtro y el espaciado se calcula mediante un intervalo constante de la frecuencia de Mel, el número de filtros puede estar entre los 20 y 40, aunque comúnmente el estándar es 26.

En la figura 21 podemos observar una gráfica con 20 filtros.

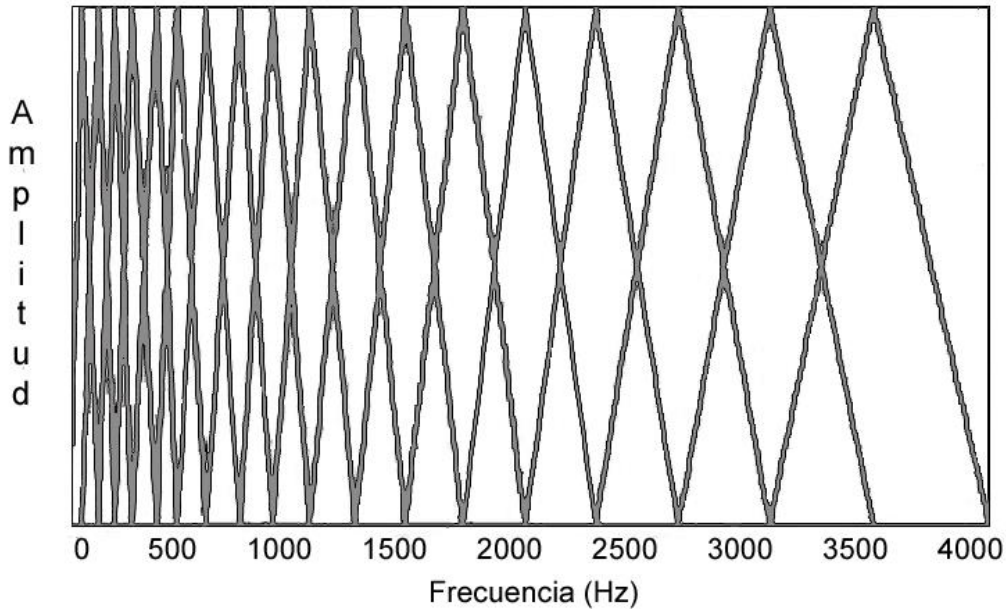

Fig. 21. Banco de filtros

Los Coeficientes Cepstrales en la Frecuencia de Mel (MFCC) son una representación definida como el cepstrum real de una señal ventaneada en tiempo corto derivada de la FFT de esa señal, la diferencia del cepstrum real es que una escala de frecuencia no lineal es usada, lo cual se aproxima al comportamiento del sistema auditivo.[23]

El significado de cepstrum, es spectrum, esto invirtiendo sus letras, en español se traduce como Cepstro y su inverso sería espectro.

El análisis Cepstro es una técnica adelantada que consiste en tomar un espectro de un espectro. Antes de calcular el cepstro, se calcula el logaritmo natural de la amplitud del espectro. El cepstro está relacionado con la función de autocorrelación; si el espectro no se hace a una escala logarítmica, el cálculo del cepstro producirá la autocorrelación. En el análisis de cepstro se trata a un espectro como si fuera una forma de onda, y se hace otro espectro a partir del primero. El

[23] H. Huang., Acero, A., H. W. Hon. Op. Cit. Pág. 314.

eje horizontal del cepstro está relacionado con el tiempo, pero no es tiempo en el sentido convencional. Se le podría llamar tiempo periódico, y de todos modos se le mide en segundos. El aspecto útil del cepstro es que extrae patrones periódicos, esos patrones que se repiten en un espectro, de la misma manera que un espectro extrae patrones periódicos de una forma de onda.[24]

2.7 Alineamiento temporal dinámico

En esta etapa se relaciona la señal de entrada con las señales guardadas en el sistema, para eso se utiliza el algoritmo DTW.

El concepto de programación dinámica, también conocido como alineamiento temporal dinámico (DTW) en reconocimiento de voz, se ha utilizado ampliamente para derivar la distorsión global entre dos muestras de voz. En estos sistemas basados en muestras, cada muestra de voz consiste en una secuencia de vectores de voz.[25]

El algoritmo DTW se usa para medir la similitud entre dos secuencias temporales, para llevar a cabo este proceso es necesario el uso de una distorsión temporal. En el DTW se comparan las distancias euclidianas entre vectores de diferentes secuencias, en este caso nuestras secuencias son la señal de voz de entrada y la previamente guardada a las que llamaremos X_n y Y_m respectivamente. Nuestras señales en este punto ya están segmentadas y enventanadas, donde m y n son los segmentos de la señal respectivamente, y estos son a su vez vectores Cepstrales desde que se aplicó la técnica de MFCC. Mientras menor distancia euclidiana entre los vectores Cepstrales, más similares serán los sonidos en estos vectores. Cuando aplicamos el DTW lo que hacemos es comparar los segmentos con menor distancia euclidiana, esto ya que al pronunciar una palabra, en distintas

[24] White, Glen. "Introducción al Análisis de Vibraciones". 2010. Pág. 90.
[25] Huang, Xuedong., Acero, Alex., Hon Hsiao, W. Op. Cit. Pág. 314.

ocasiones puede haber diferencias, como la velocidad de la pronunciación, o factores como el volumen o nerviosismo de diferentes locutores. De esta manera con el DTW comparamos los vectores fonéticamente homólogos.

En la figura 22 podemos observar la comparación entre dos secuencias de voz, y el alineamiento entre las mismas.

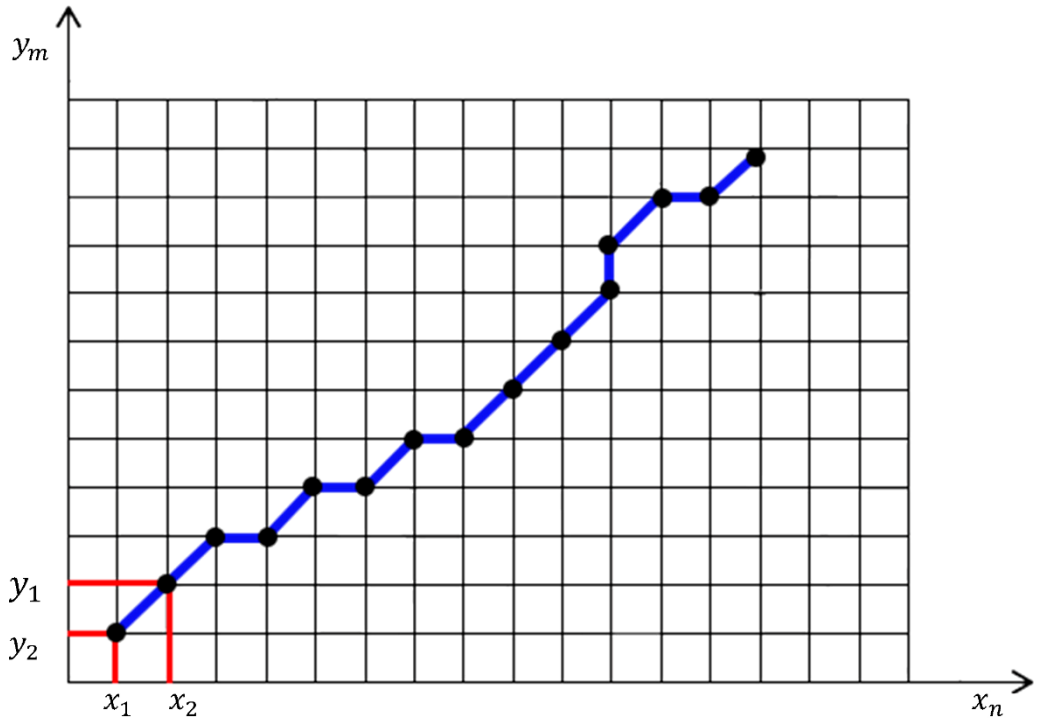

Fig. 22. Ilustración del DTW

Para entender mejor el funcionamiento del algoritmo DTW, en la figura 23 se muestran 2 señales segmentadas, $a(t)$ y $b(t)$, donde t=tiempo, al aplicar este algoritmo los segmentos se alinean en el tiempo y se comparan con los de menor distancia euclidiana, los segmentos comparados en este caso han sido sombreados.

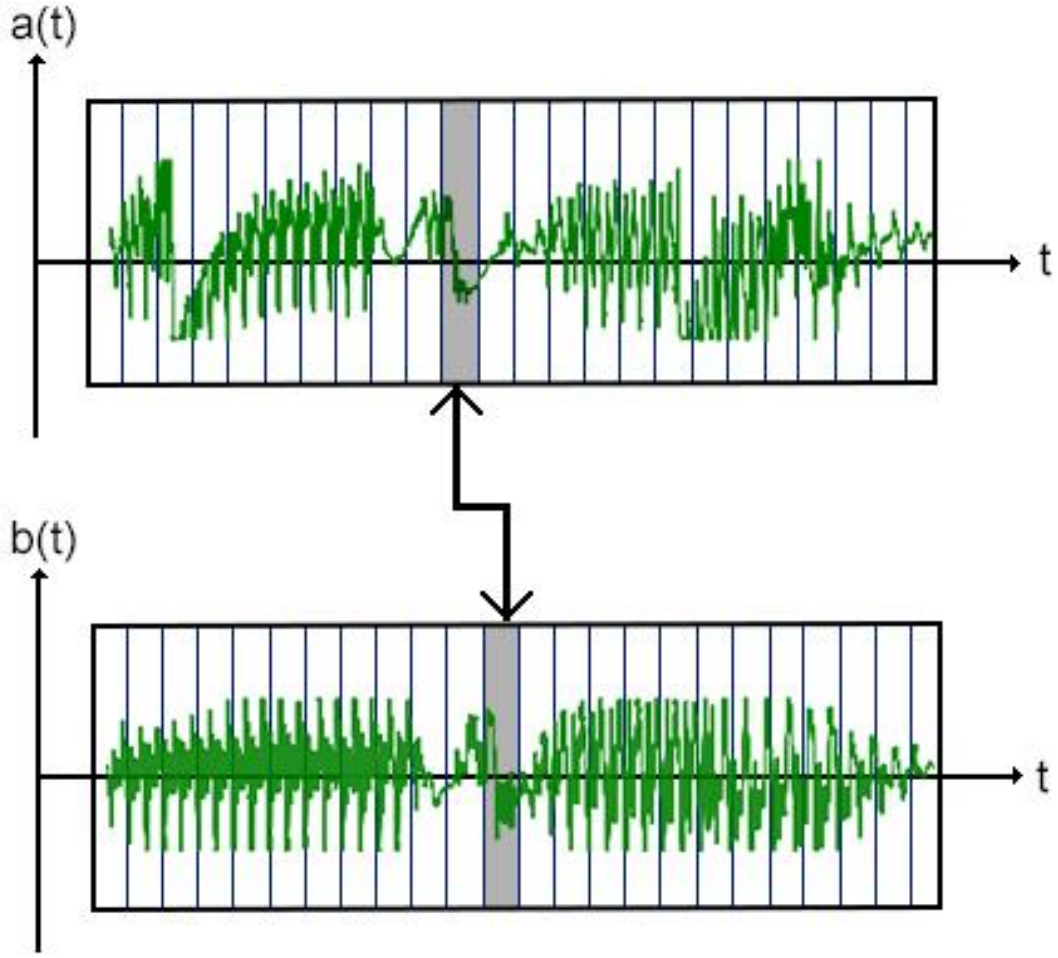

Fig. 23. Función segmentada y alineada en el tiempo normalizado

El habla es un proceso variable en el tiempo, en el cual la duración de una palabra y sus sub-palabras varían aleatoriamente. Por lo tanto, se requiere un método para encontrar la mejor alineación de tiempo entre una secuencia de características vectoriales que representan una palabra hablada y los candidatos modelo. La mejor alineación de tiempo entre dos secuencias de vectores se puede definir como la alineación con la distancia euclidiana mínima. Para el reconocimiento de palabras aisladas, el método de alineación de tiempo utilizado es el alineamiento temporal dinámico (DTW).[26]

[26] Saeed V. Vaseghi. Op. Cit. Pág 523

2.8 Reconocimiento de la palabra y acción a realizar

En el paso anterior se determina si la palabra de entrada es parecida con alguna de las previamente guardadas, esto por medio de la distancia euclidiana, en este punto se determina si la palabra coincide o no con otra almacenada en la base de datos, posteriormente si la palabra es reconocida, el sistema produce una señal de salida para realizar una acción. La tabla 1 muestra la relación entre la palabra reconocida y la acción a realizar.

Palabra (Comando) de entrada	**Salida**	**Acción a realizar**
Arriba	GPIO 5	Prender LED
Abajo	GPIO 6	Prender LED
Izquierda	GPIO 13	Prender LED
Derecha	GPIO 19	Prender LED
Avanzar	GPIO 26	Prender LED
Atrás	GPIO 14	Prender LED
Parar	GPIO 15	Prender LED

Tabla. 1. Relación entra la entrada y la acción a realizar

2.9 Base de datos

Se ha desarrollado una base de datos local, es decir que sólo se encuentra en el sistema empotrado que se ha utilizado, con la finalidad de almacenar señales de entrada, que posteriormente servirán para realizar comparaciones con las nuevas señales de entrada.

Una base de datos es una colección de datos, que generalmente describe las actividades de una o más organizaciones relacionadas.[27]

[27] Raghu Ramakrishnan, Johannes Gehrke. "Database Management Systems". 3ª Edición. 2003. Pág. 1

Capítulo 3 Implementación del algoritmo de RAH

El Hardware necesario en el sistema empotrado para poder llevar a cabo la implementación del algoritmo de RAH, consiste en una tarjeta micro SD de 16Gb, previamente cargada con el sistema operativo Ubuntu MATE, una tarjeta de audio USB manhattan que funciona como CAD y CDA, un micrófono para capturar las señales de audio, disipadores de calor para que el sistema pueda funcionar sin sobre calentarse, un push button y una carcasa oficial de Raspberry pi. Todos los elementos mencionados se muestran en la figura 24.

Fig. 24. Elementos ocupados en el sistema empotrado

Un sistema operativo es un software que gestiona el hardware de la computadora. El hardware debe proporcionar los mecanismos apropiados para

asegurar el correcto funcionamiento del sistema informático e impedir que los programas de usuario interfieran con el apropiado funcionamiento del sistema.[28]

3.1 Configuración

La tarjeta de sonido USB se ha configurado para ser reconocida por el sistema operativo como la tarjeta principal de audio. Los pasos para realizar la misma, comienzan por tener acceso como súper usuario desde la consola, lo cual se hizo con el siguiente comando:

$$sudo\ su$$

Posteriormente se debió modificar el siguiente archivo con el comando $nano$, localizado en la dirección

$$/etc/modprobe.d/raspi - blacklist.cof$$

Y se agregó la tarjeta default del sistema empotrado, con la siguiente instrucción:

$$blacklist\ snd\_bcm2835$$

Como se muestra en la figura 25:

Fig. 25. Archivo modificado "blacklist.conf"

[28] Silberschatz, Abraham. "Fundamentos de sistemas". 7ª Edición, 2006. Pág. 2.

Esta línea permitió bloquear la tarjeta default del sistema.

El siguiente paso es agregar la tarjeta externa de audio USB, modificando el archivo ubicado en la siguiente dirección:

$$/usr/share/alsa/alsa.conf$$

Dentro del archivo $alsa.conf$ se buscó la siguiente línea de instrucción:

$$options\ snd-usb-audio\ index-2$$

Y se cambió el índice -2 por el índice 0, como se muestra a continuación:

$$options\ snd-usb-audio\ index\ 0$$

Al terminar este proceso la tarjeta de audio USB Manhattan queda definida como la tarjeta principal de audio del sistema empotrado.

Para la implementación del algoritmo de RAH se ha utilizado el lenguaje de programación Python.

Python es un lenguaje de programación de computadoras que nació en 1991 y que ha ido ganando adeptos por dos razones principales. La primera es que es un lenguaje de alto nivel muy fácil de usar y la segunda es que es de código libre.[29]

Una de las grandes ventajas de Python es que cuenta con una gran variedad de librerías que le da al usuario una gran cantidad de funcionalidades para realizar múltiples tareas en esta plataforma de programación.

[29] Cervantes, Ofelia D., Báez López, David., Arizaga Silva, Antonio., Castillo Juárez, Esteban. "Python con aplicaciones a las matemáticas, ingeniería y finanzas". 1ª Edición, 2017. Pág. 14.

3.2 Instalación de Software

La versión usada en este proyecto es Python 3, la cual es más actual de la plataforma, el comando para instalar esta versión es el siguiente:

$$sudo\ apt - get\ install\ python3$$

Probamos el entorno de programación Python 3 desde la consola de Ubuntu MATE, para lo cual sólo damos la instrucción $nano$ más el nombre del archivo que deseamos crear y la extensión, en este caso. py, por ejemplo “holamundo.py”. En la figura 26 se muestra el entorno de programación Python en modo consola.

$$nano\ Nombre\_del\_archivo.py$$

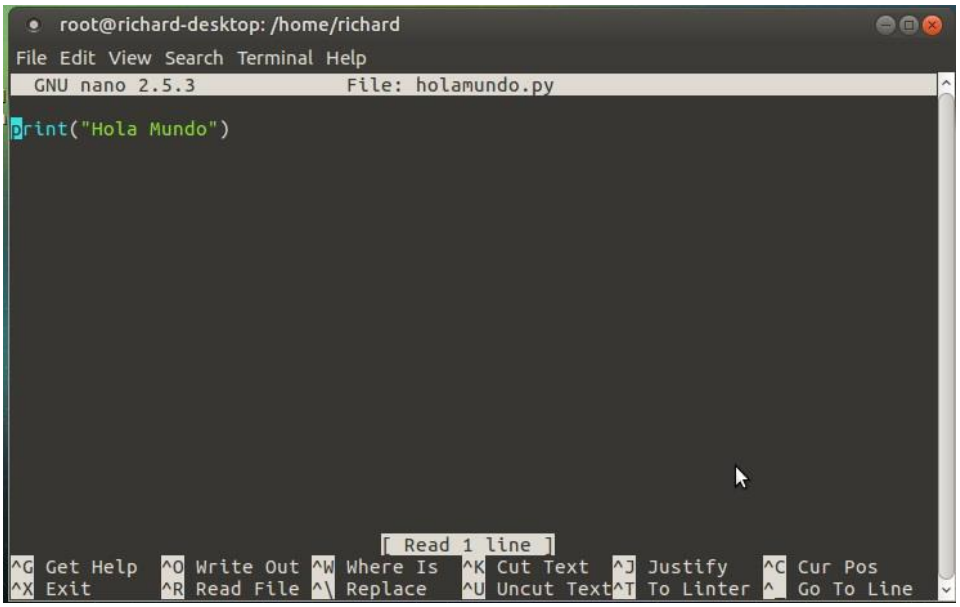

Fig. 26. Ejemplo de programa en Python en consola.

Posteriormente se instalaron algunas librerías para lograr nuestros objetivos, todo el proceso se lleva a cabo desde la consola del sistema, a continuación, se muestra cómo se llevó a cabo dicha instalación:

$$python\ -m\ pip\ install\ pyaudio$$
$$sudo\ pip3\ install\ numpy\ spicy\ matplotlib$$

Para que las librerías sean reconocidas correctamente, se agregaron las correspondientes a la versión de Python utilizada. Lo cual hacemos con los siguientes comandos desde consola:

$$python3$$
$$import\ numpy, scipy, matplotlib$$
$$quit()$$

Este proceso se muestra en la figura 27:

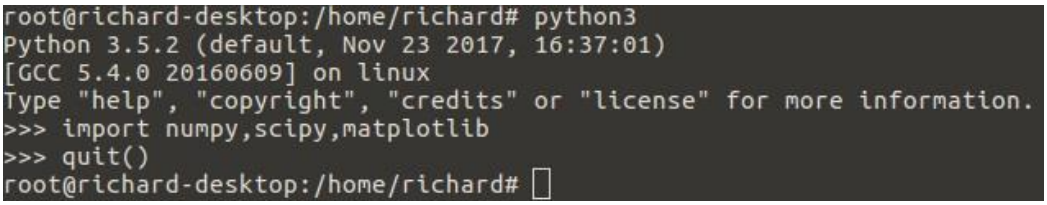

Fig. 27. Proceso para agregar las librerías Numpy, Scipy, Matplotlib a Python 3

Muchos de los sistemas de RAH son implementados en Matlab, en este trabajo se presenta una alternativa utilizando Python con las librearías Matplotlib y Numpy, que en conjunto nos permitirán realizar procesos matemáticos complejos, y graficar en 2D cuando sea necesario.

La principal diferencia entre Matlab y nuestro lenguaje de programación equipado con las librerías antes mencionadas, es que, Matlab es un software de paga, el cual no está disponible para nuestro sistema empotrado y al ser un Software con entorno gráfico, consume más recursos que Python, ya que se utiliza desde consola. Además, maneja su propio lenguaje de programación (Lenguaje

M) lo que provocaría que nuestro sistema resulte inservible en otros dispositivos electrónicos que no cuenten con el Software de Matlab previamente instalado.

3.3 Implementación del algoritmo en el sistema empotrado

Procederemos a utilizar la información que se mostró en el Capítulo 2 para implementar la propuesta de algoritmo en Python de esta tesis y poder realizar las pruebas correspondientes para observar y medir su comportamiento.

3.3.1 Captura de audio

Se utilizó un push button conectado a los pines de entrada/salida de propósito general (General Purpose Input/Output; GPIO) del sistema empotrado, para que al momento de presionar y mantener presionado este botón se inicie la grabación por medio de la tarjeta de audio USB previamente configurada, la conexión se muestra en la figura 28.

Como el acrónimo sugiere los pines GPIO pueden aceptar comandos tanto de entrada como de salida controlados por una variedad de lenguajes ejecutados en la Raspberry pi.[30]

[30] A. K. Dennis. Op. Cit. Pág. 5.

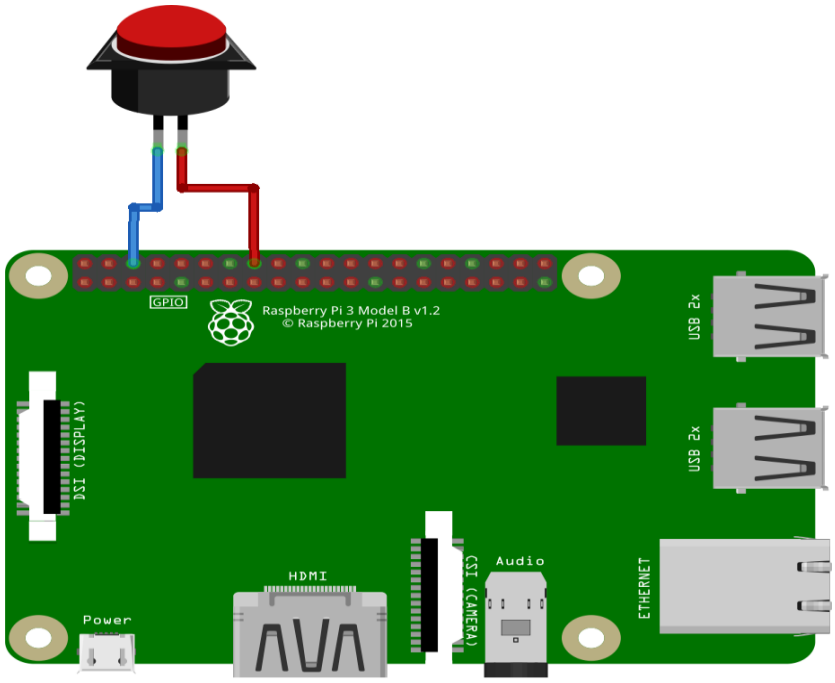

Fig. 28. Conexión del Push Button en la Raspbery pi.

Se agregan las librerías que usaremos al código, como se muestra en la figura 29

```
import wave
import RPI.GPIO as GPIO
import time
import matplotlib.pyplot as plt
import numpy as np
import scipy,fftpack
import sys
```

Fig. 29. Librerías agregadas al programa de Python

Para empezar el código, declaramos un pin GPIO de la Raspberry como entrada, el push button se conectó a éste y a un pin de tierra, además declaramos los valores de captura que definimos en el capítulo anterior, esto lo podemos observar en la figura 30.

```
GPIO.setmode(GPIO.BCM)
GPIO.setup(23, GPIO.IN, pull_up_down=GPIO.PUD_UP)
FORMAT = pyaudio.paInt16
CHANNELS = 1
RATE = 11025
```

Fig. 30. Código de Python con la declaración de GPIO y variables.

Para utilizar el botón, se creó un ciclo para realizar la acción de grabar, esto se puede observar en la figura 31. Dicha acción se realiza mientras el botón este presionado, en el momento en el cual éste se deja de presionar, la grabación se termina y se procede a guardar la misma en un archivo de audio WMA.

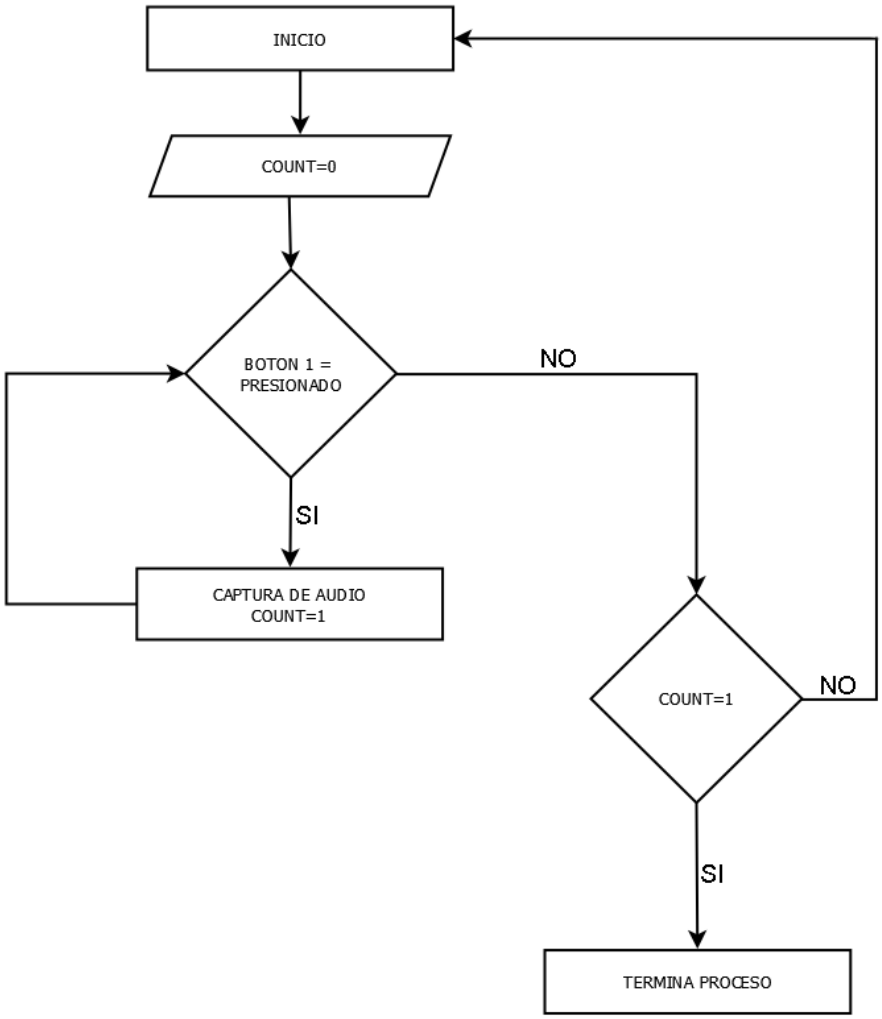

Fig. 31. Diagrama a bloques del ciclo de grabación

Cuando el ciclo del botón termina, y el archivo ha sido guardado exitosamente, ya es posible manipular el audio a nuestra conveniencia.

Con la orden $plt.plot$ podemos graficar la señal de entrada que se ha grabado al presionar el botón, como se muestra en la figura 32.

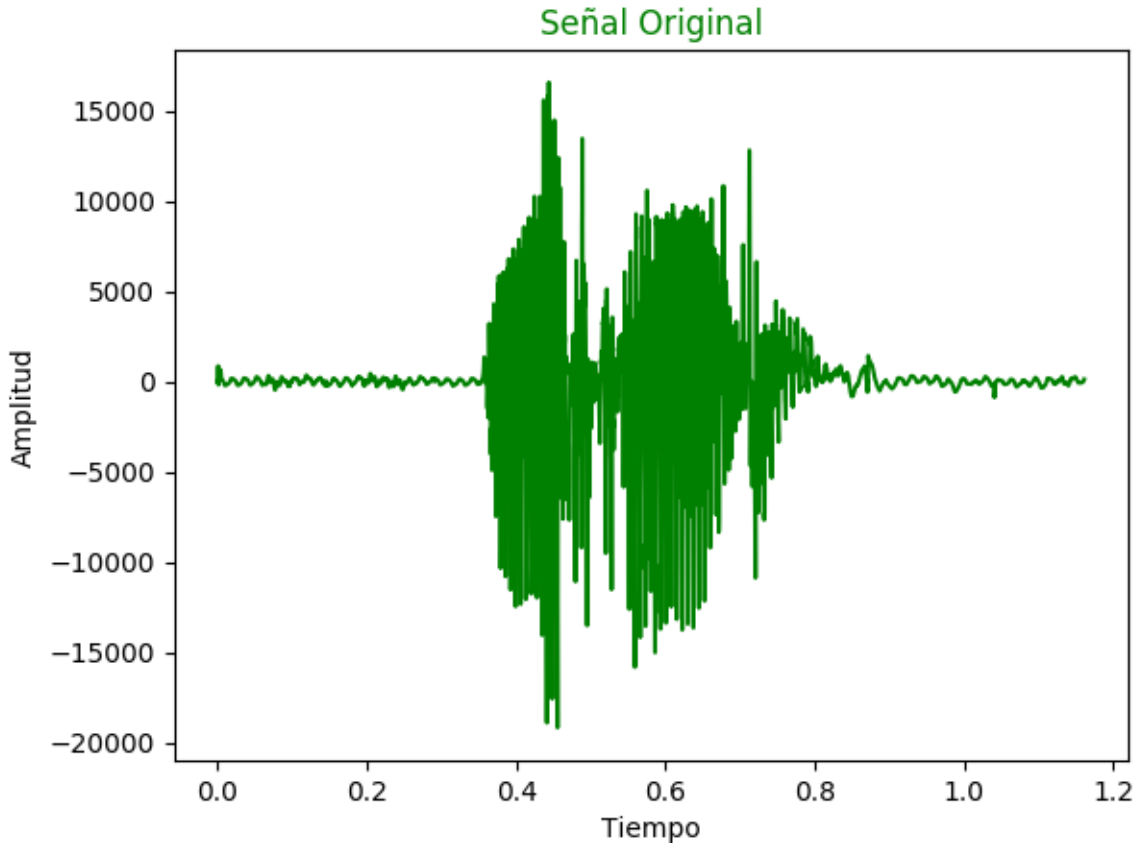

Fig. 32. Señal original graficada

3.3.2 Corte de silencio

El siguiente paso del algoritmo es el corte de silencio, como se presento en el Capitulo anterior, es necesario aplicar un filtro pasa altas para eliminar frecuencias por encima de una F_c, esto se aplica a la señal original, y el resultado es una señal donde las frecuencia inecesarias han sido eliminadas.

Una vez realizado el proceso, se graficará para comprobar la señal de salida, como se muestra en la figura 33.

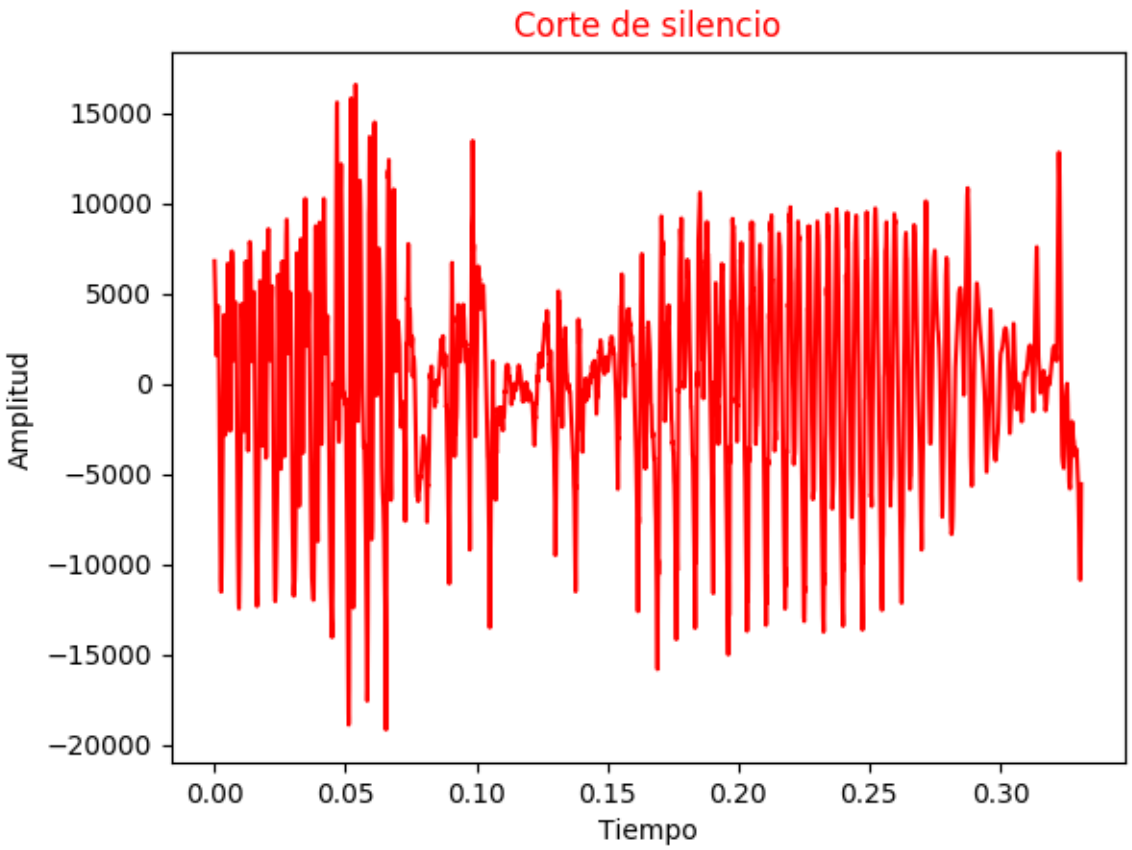

Fig. 33. Señal con corte de silencio aplicado.

La señales de las figuras 32 y 33, son las mismas, pero en la última se ha eliminado exitosamente el silencio, para comparar ambas, se sobrepone una encima de la otra como se muestra en la figura 34.

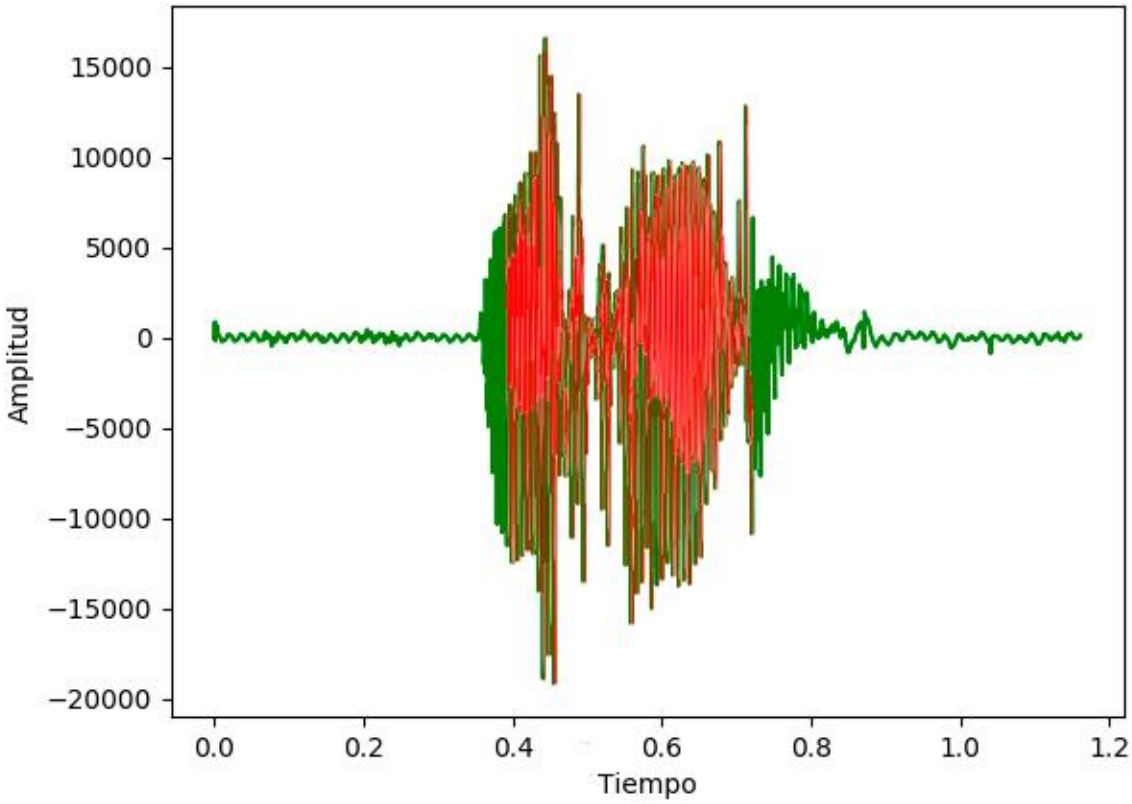

Fig. 34. Señal de corte de silencio sobrepuesta a la señal original.

3.3.3 Preénfasis

Aplicamos el filtro de preénfasis usando los valores predeterminados, que son 0.95 o 0.97, en este caso usaremos el segundo, y lo declararemos en una variable como $\text{pre\_emphasis} = 0.97$ posteriormente se aplicó a la señal.

Una vez aplicado el filtro de pre-enfasis se grafica nuevamente la señal, como se muestra en la figura 35.

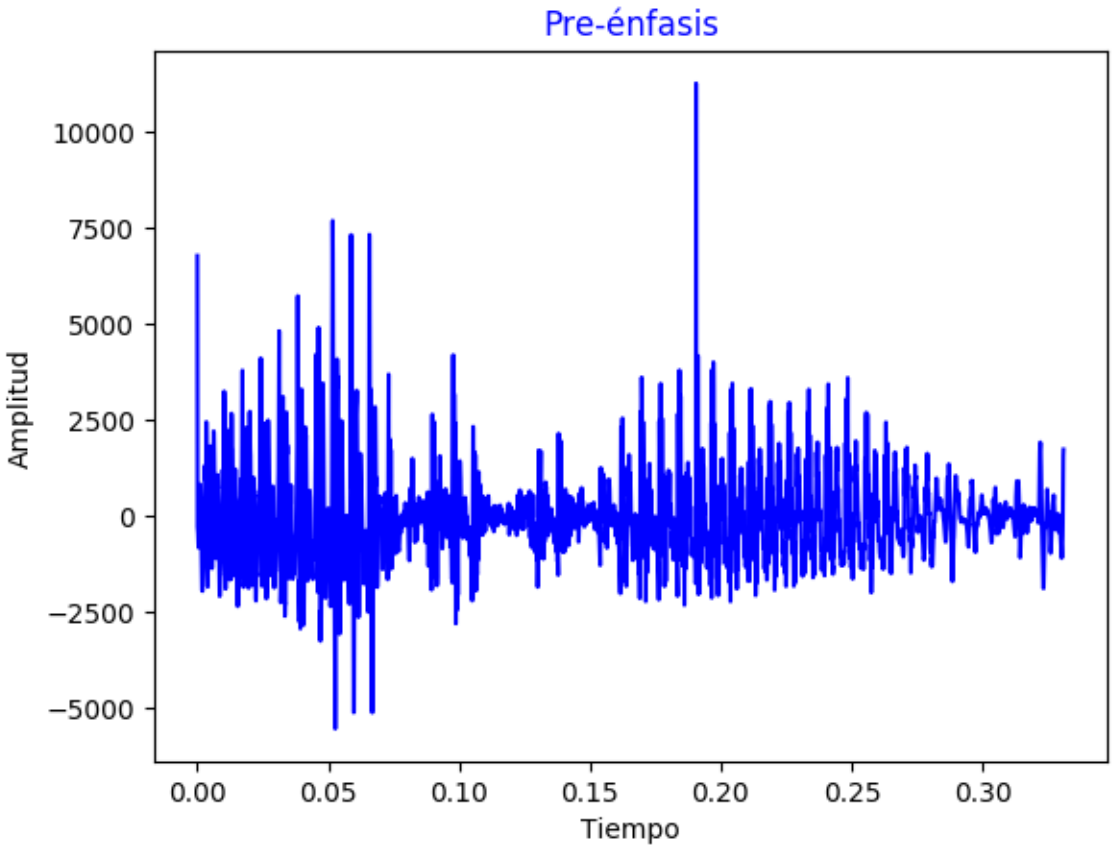

Fig. 35. Señal con filtro de preénfasis aplicado

3.3.4 Segmentación de la seañal y ventaneo de hamming

La segmentación de la señal utilizada fue de 25ms ya que con esta la señal será estacionaria, para esto sólo se creo una variable con el valor ya establecido y posteriormente se aplicó la venta de Hamming a cada segmento de la señal, en la figura 36 se puede observar un segmento con el ventaneo aplicado.

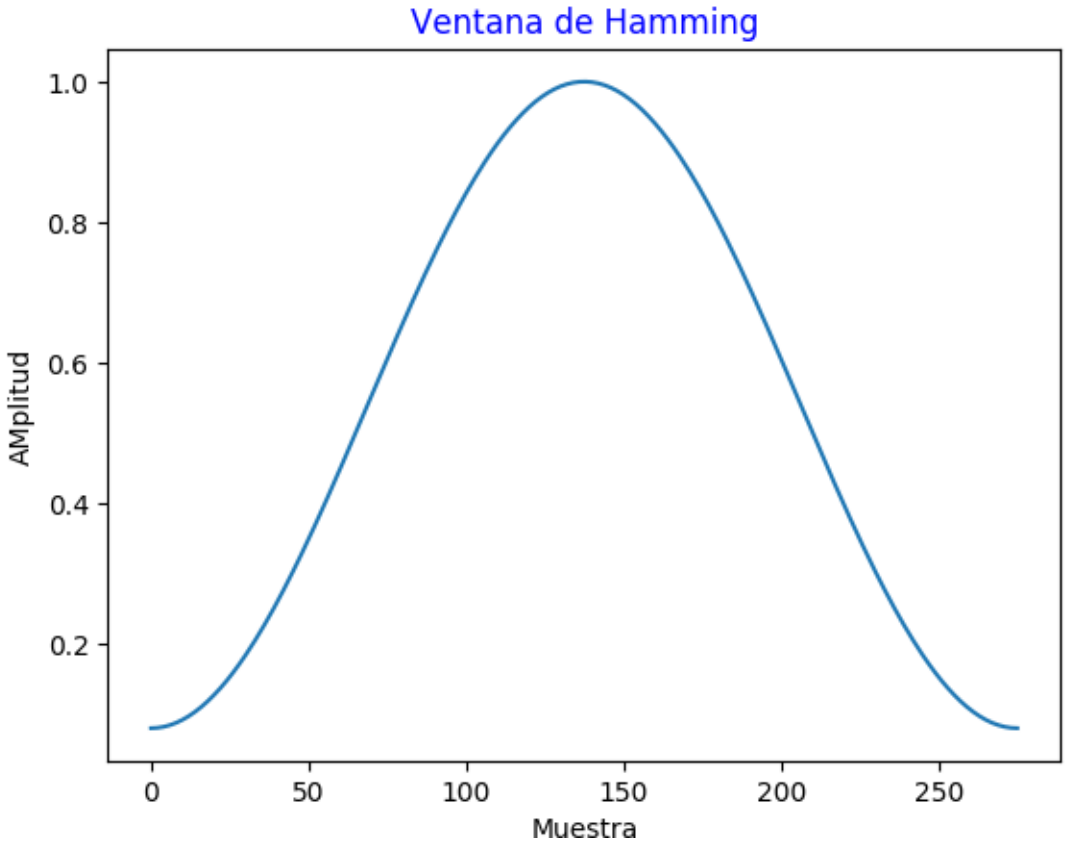

Fig. 36. Segmento de la señal con la venta de Hamming aplicada

3.3.5 STFT, Filter Banks y MFCC

1. Se aplica la STFT a cada uno de los segmentos de la señal ventaneada.
2. Se aplican los filtros triangulares, esto para calcular los bancos de filtros, en este caso usamos 40 filtros.
3. Finalmente aplicamos los MFCC a la señal, utilizamos los coeficientes Cepstrales 2-13.

En la figura 37 observamos el resultado de graficar los MFCC.

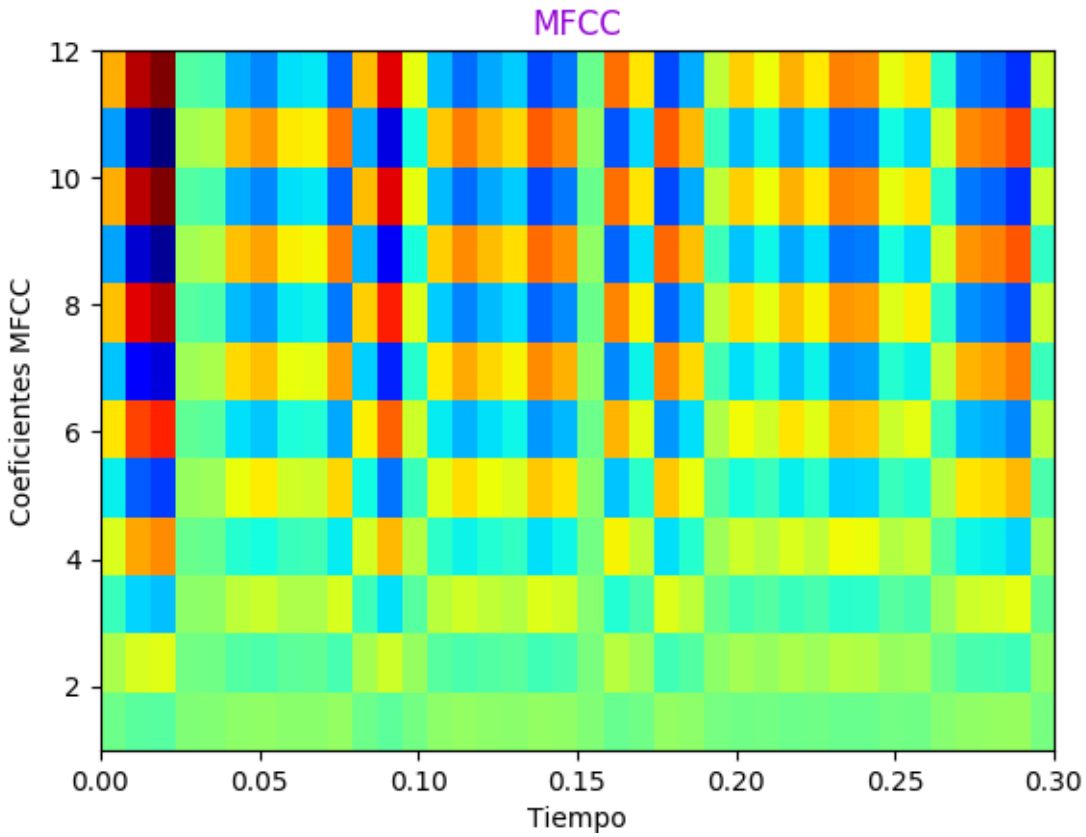

Fig. 37. MFCC de la señal

3.3.6 DTW

Una vez que el algoritmo ha realizado todas las etapas previas, se procede a determinar si la palabra (señal) de entrada, coincide con alguna de las previamente guardadas en la base de datos, esto se hace con el algoritmo DTW, como se explicó en el Capítulo anterior, éste compara las distancias euclidianas de cada vector. En la figura 38, se muestra la comparación de dos señales diferentes, pero con la misma palabra pronunciada.

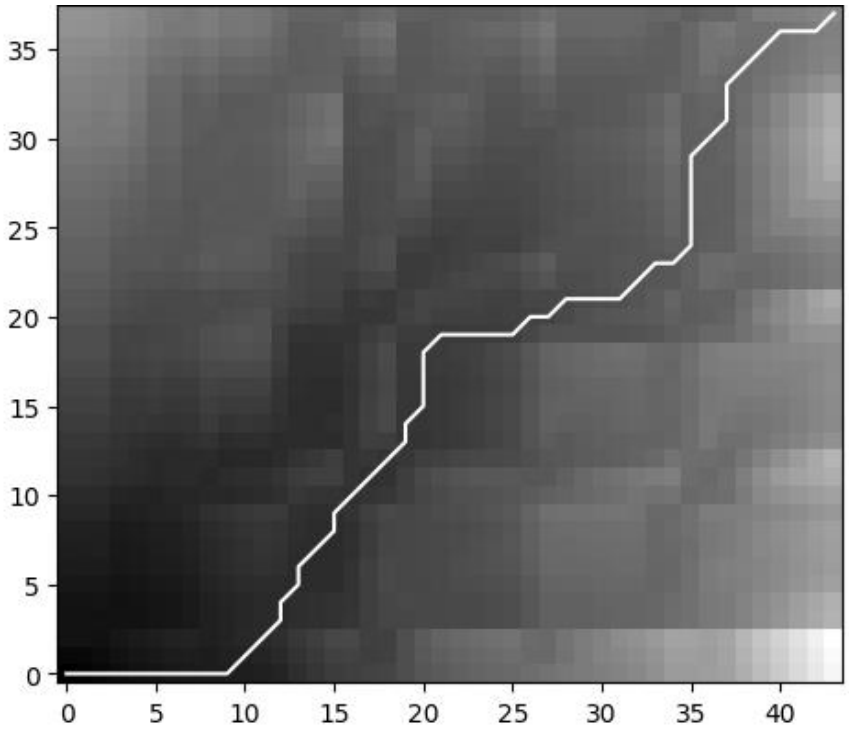

Fig. 38. DTW de dos señales

3.3.7 Base de datos

Para la base de datos se usaron un total de 140 archivos, teniendo 20 repeticiones por cada palabra diferente, con 10 diferentes usuarios.

En la imagen 39 se muestra el diagrama entidad relación de la base de datos utilizada.

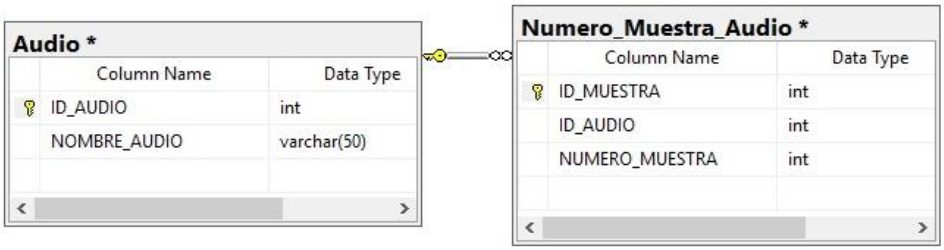

Fig. 39. Diagrama entidad-relación de la base de datos.

Capítulo 4 Pruebas de funcionamiento

Después de implementar el algoritmo, se realizaron las pruebas de funcionamiento en un ambiente controlado de laboratorio, las cuales consistieron en que un usuario, pronuncie los comandos previamente definidos, para probar el sistema de RAH.

Para determinar si una palabra coincide con otra previamente guardada en la base de datos, se revisaron las distancias euclidianas que hay entre ellas. Con esto se estableció un criterio para determinar la distancia máxima para que una palabra sea reconocida.

Para estas pruebas se han conectado 7 LED´s de diferentes colores al sistema empotrado, los cuales representan las funciones que realiza el BRCL. Los LED's conectados al sistema empotrado se muestran en la figura 40.

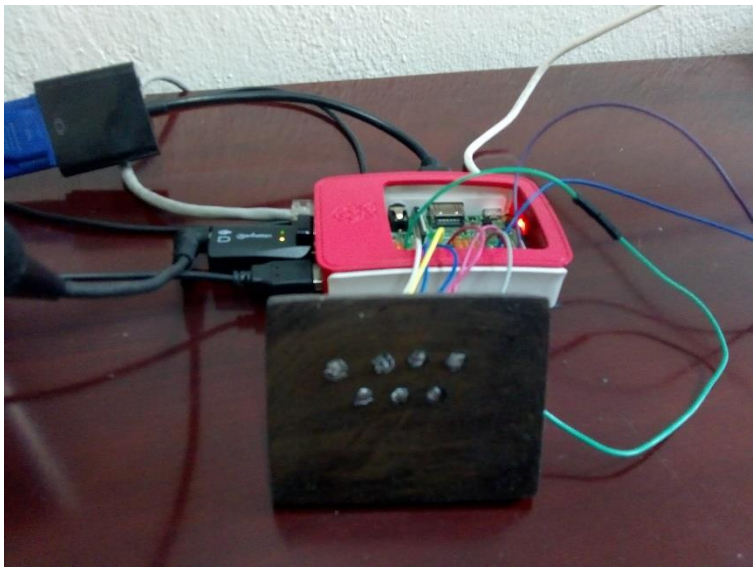

Fig. 40. Sistema empotrado con LED´s conectados a las salidas GPIO.

En la tabla 2 se muestra cada palabra de entrada asociada a cada GPIO de salida y cada LED con su respectivo color.

Palabra (Comando) de entrada	**Salida**	**Color de LED**
Arriba	GPIO 5	 Naranja
Abajo	GPIO 6	 Blanco
Izquierda	GPIO 13	 Violeta
Derecha	GPIO 19	 Naranja
Avanzar	GPIO 26	 Rojo
Atrás	GPIO 14	 Verde

Parar	GPIO 15	 Azul

Tabla. 2. Relación entre señales de entrada y sus salidas a cada LED y su respectivo color

Metodología del experimento:

1. El usuario se coloca a una distancia aproximada de 10 cm del micrófono.
2. Se presiona el push button para iniciar el sistema.
3. Se pronuncia la palabra de entrada.
4. Al terminar la pronunciación, se deja de presionar el push button.
5. Se espera a que el sistema procese la señal de entrada (palabra) y realice una acción (prende el LED predeterminado).

El proceso de funcionamiento mencionado se ilustra en la imagen 41.

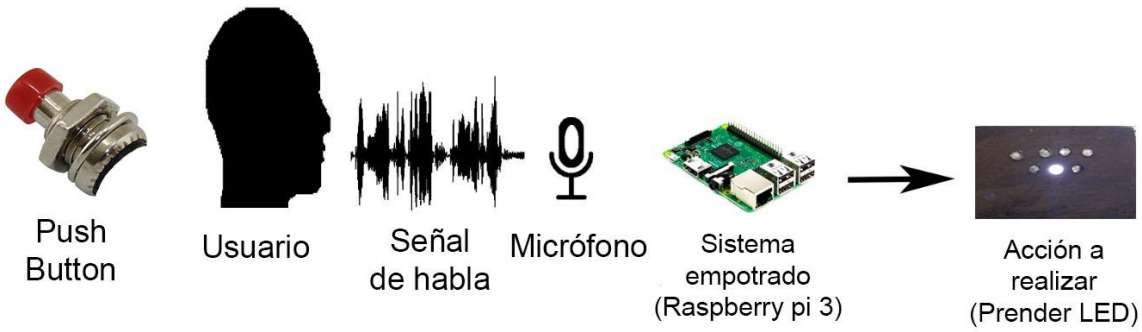

Fig. 41. Ilustración del funcionamiento del sistema de RAH

Resultados

La tabla 3, muestra el número de veces que se probó el sistema y el porcentaje de aciertos obtenidos. Las pruebas se realizaron con 10 distintos usuarios.

El algoritmo tarda un promedio de 2.19 segundos para determinar si la palabra coincide con una de las previamente guardadas en la base de datos.

Palabra (Comando) de entrada	**Numero de pruebas por usuario**	**Porcentaje de aciertos**
Arriba	10	90%
Abajo	10	91%
Izquierda	10	84%
Derecha	10	87%
Avanzar	10	90%
Atrás	10	90%
Parar	10	93%

Tabla. 3. Resultados de la prueba del sistema

Conclusiones y Recomendaciones (Trabajos futuros)

Se logró cumplir el objetivo general de diseñar e implementar un sistema de control por voz, por medio de distintas técnicas, métodos y filtros matemáticos los cuales fueron posibles adaptar gracias a la investigación de diferentes algoritmos de RAH

Los objetivos específicos se cumplieron:

- Se identificaron distintos tipos de artículos y trabajos científicos donde se abordaba el tema de implementación de algoritmos de RAH.
- Se diseñó un algoritmo propio de RAH basado en diferentes técnicas, métodos y filtros matemáticos.
- Con la definición del problema fue posible establecer el numero de comandos que es necesario identificar por el sistema.
- Es posible manipular un BCRL por medio de comandos de voz, esto se simulo por medio de leds, los cuales representan cada una de las funciones que el BCRL es capaz de realizar.
- Se evaluó el funcionamiento del sistema de RAH, por medio de las pruebas de funcionamiento.
- Los avances y resultados son reportados en el presente trabajo de tesis.

Las hipótesis fueron demostradas en virtud de que:

- Es posible manipular un BCRL por medio de un sistema de control utilizando comandos de voz. Esto es viable porque al simularlo, el sistema responde de manera correcta en un 89.2% de las pruebas realizadas.

- El algoritmo diseñado para el sistema de control por voz puede ser implementado en cualquier sistema empotrado. Esto es verídico, sólo es necesario implementar el algoritmo diseñado en el lenguaje de programación requerido por el sistema empotrado que sea utilizado.

Se logró implementar el algoritmo en un sistema empotrado, para esto se utilizó un lenguaje de programación el cual es reconocido por dicho sistema, se concluye que, siguiendo el algoritmo propuesto paso a paso no existe limitación alguna para adaptarlo al lenguaje de programación utilizado.

Se concluye que el sistema es funcional y soporta distintas señales de voz de entrada.

Como trabajo futuro se propone optimizar el algoritmo para que la ejecución del mismo se haga en el menor tiempo posible, además de esto se requiere realizar un sistema empotrado propio que se adapte a las necesidades específicas que se requieren.

Glosario

Amplitud: (En audio) es el grado de movimiento de las moléculas de aire en la onda, que corresponde a la intensidad del enrarecimiento y compresión que la acompañan.

Convolución: Es un operador matemático que transforma dos funciones f y g en una tercera función que en cierto sentido representa la magnitud en la que se superponen f y una versión trasladada e invertida de g.

Decibel: El término decibel se deriva del hecho de que la potencia y los niveles de audio guardan una relación logarítmica. Esto es, un incremento del nivel de potencia de, por ejemplo, 4 W a 16 W no aumenta el nivel de audio por un factor de $16/4 = 4$, sino por un factor de 2, como se deduce de la potencia de 4 de la manera siguiente: $4^2 = 16$

Distancia Euclidiana: En matemáticas, la distancia euclidiana o euclídea es la distancia "ordinaria" (que se mediría con una regla) entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras.

Diagrama a bloques: Un diagrama en el que las unidades de sistemas esenciales se dibujan como bloques, y su relación entre sí se indica mediante líneas apropiadamente conectadas. La ruta de la señal o energía puede ser indicada por líneas o flechas.

Enventanamiento: Proceso en el cual se le aplica una función ventana (Ejemplo: Hamming, Hanning, rectangular, etc.) a una señal estacionaria.

Escala Mel: Es una escala perceptual de tonos juzgados por los oyentes para ser iguales en distancia el uno del otro. El punto de referencia entre esta escala y la medición de frecuencia normal se define asignando un tono perceptual de 1000 Mels a un tono de 1000 Hz, 40 dB por encima del umbral del oyente. Por encima de alrededor de 500 Hz.

Espectrograma: Representación gráfica de sonidos hecha en una máquina en términos de sus componentes de frecuencia. El tiempo se muestra en el eje horizontal, la frecuencia en el eje vertical y la intensidad en la oscuridad de la marca.

Ganancia: Cualquier incremento en poder cuando una señal es transmitida de un punto a otro. Usualmente se expresa en decibeles.

Hardware: En informática se refiere a las partes físicas o los componentes de una computadora, como la unidad de procesamiento central, el monitor, el teclado, el almacenamiento de datos de la computadora, la tarjeta gráfica, la tarjeta de sonido, los parlantes y la placa base

Hertz: La unidad de frecuencia. Un Hert es igual a un ciclo por segundo. Su abreviatura es HZ, nombrado así por H. R. Hertz, un físico alemán del siglo XIX.

LED: Diodo emisor de luz o light-emitting diode (LED por sus siglas en ingles), es un dispositivo semiconductor que emite luz visible cuando se desplaza hacia adelante.

Librería: Es una colección de implementaciones de comportamiento, escrita en términos de un lenguaje, que tiene una interfaz bien definida mediante la cual se invoca el comportamiento. Por ejemplo, las personas que desean escribir un

programa de alto nivel pueden usar una biblioteca para realizar llamadas al sistema en lugar de implementar esas llamadas al sistema una y otra vez.

Línea de comando: Un único comando o instrucción que dirige una computadora para resolver un problema, generalmente escrito en una línea.

Matplotlib: Es una biblioteca de graficación en 2D de Python que produce figuras de calidad, en una variedad de formatos “hardcopy” y entornos interactivos, es multiplataforma y para la gráfica simple produce una interfaz de tipo MATLAB

Nano: (Oficialmente GNU nano) es un editor de texto para sistemas informáticos de tipo Unix o entornos operativos que utilizan una interfaz de línea de comando.

Numpy: Es una extensión matemática de Matplotlib, la cual agrega mayor soporte para vectores y matrices, constituyendo una biblioteca de funciones matemáticas de alto nivel para operar con esos vectores o matrices

Push Button: Es un dispositivo utilizado para realizar cierta función. Los botones son por lo general activados, al ser pulsados con un dedo. Permiten el flujo de corriente mientras son accionados. Cuando ya no se presiona sobre él vuelve a su posición de reposo.

Python: Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y en menor medida, programación funcional. Es un lenguaje interpretado, usa tipado dinámico y es multiplataforma.

Software: Es un término genérico que se refiere a la colección de datos o instrucciones de máquina que le dicen a una computadora como trabajar, en contraste con el Hardware físico desde el cual el sistema es construido, que de hecho es el que realiza el trabajo. En ciencias de la computación e ingeniería de

software, el software es toda la información procesada por los sistemas computacionales, programas y datos. El Software incluye programas informáticos, librerías y datos no ejecutables relacionados, como documentación en línea o medios digitales.

System on a Chip (SoC): Es un circuito integrado (también conocido como "IC" o "chip") que integra todos los componentes de una computadora u otros sistemas electrónicos. Estos componentes generalmente incluyen una unidad de procesamiento central (CPU), memoria, puertos de entrada / salida y almacenamiento secundario, todo en un único sustrato.

Filtro digital: Es un filtro que opera con señales digitales, como el sonido representado dentro de una computadora. Es un cálculo que toma una secuencia de números (la señal de entrada) y produce una nueva secuencia de números (la señal de salida filtrada). Tipo de circuito que deja pasar ciertas frecuencias y rechaza todas las demás.

Filtro pasa-altas: Tipo de filtro que deja pasar todas las frecuencias ubicadas por encima de una frecuencia crítica y rechaza todas las frecuencias localizadas por debajo de dicha frecuencia de corte.

Frecuencia: Medida de la razón de cambio de una función periódica; es el número de ciclos completados en 1 segundo. La unidad de frecuencia es el Hertz.

Frecuencia de corte: Frecuencia a la cual el voltaje de salida de un filtro es del 70.7% del voltaje de salida máximo.

Frecuencia de Muestreo: La velocidad con la cual la muestra analógica es medida o mostrada por cada segundo

Función de transferencia: Es la relación entre una función forzada y una función de excitación (o entre una salida y una entrada) dependiente de la frecuencia.
Scipy: Es un ecosistema basado en Python de software de código libre para matemáticas, que incluye múltiples paquetes, como pueden ser Matplotlib, Numpy, Ipython, Simpy, entre otros.

Secure Digital (SD): Es un formato de tarjeta de memoria. Se utiliza en dispositivos portátiles, son el formato de tarjetas de memoria más utilizado en la actualidad.

Transformada de Fourier: Es una operación matemática que transforma una señal de dominio de tiempo a dominio de frecuencia y viceversa.

Universal Serial Bus (USB): Se trata de un concepto de la informática para nombrar al puerto que permite conectar periféricos a una computadora

WAV: Es un formato de audio digital normalmente sin compresión de datos desarrollado. Creado por Microsoft e IBM se utiliza para almacenar sonidos en el PC, admite archivos mono y estéreo a diversas resoluciones y velocidades de muestreo, su extensión es .wav

Bibliografía:

Balosa, J.; Crespo, F.J.; Bariiga, A. “Sistema empotrado de reconocimiento de voz sobre FPGA”, Universidad de Sevilla. 2012

Bishop, Christopher M. “Pattern Recognition and Machine Learning”, Ed. Springer, Singapore, 2007.

Bojan K.; Simon J.; Szakáll T.; Tibor S. “Mobile robot controlled by voice”, IEEE, 2007

Boylestad, Robert L.; Nashelsky, Louis. “Electrónica: Teoría de circuitos y dispositivos electrónicos”, Ed. Prentice Hall, México 2009

Carlson A. Bruce.; Crilly, Paul B. “Communication systems - An Introduction to Signals and Noise in Electrical Communication”, Ed. McGraw-Hill, New York, NY 2010

Dennis, Andrew K. “Raspberry Pi Computer Architecture Essentials” Ed. Packt Publishing, UK 2016.

Donat, Wolfram. “Learn Raspberry Pi Programming with Python”, Ed. Technology in action. 2014.

Floyd, Thomas L. “Principios de circuitos eléctricos”, Ed. Pearson Educación, México, 2007.

Gibilisco, Stan. “The Illustrated Dictionary of Electronics”, Ed. McGraw-Hill, United States of America, 2001.

Hong, Q.; Zhang, C.; Chen X.; Chen Y.; Xia.Yang C. “Embedded Speech Recognition System for Intelligent Robot”, IEEE. 2007.

Huang, G. S., Yang S. “The ASR Approach Based on Embedded System for Meal Service Robot”, IEEE, 2012.

Huang, Xuedong.; Acero, Alejandro.; Hon Hsiao, Wuen. “Spoken Language Processing: a guide to theory, algorithm, and system development”, Ed. Prentice Hall, United States of America 2001.

Kuo, Benjamin. “Sistemas automáticos de control”, Ed. Prentice Hall, México 1996.

Lyons, Richard G. “Understanding Digital Signal Processing”, Ed. Pearson Education, U.S. 2011.

Madisetti V.K.; Williams, D. B., “Digital Signal Processing Handbook” Ed. CRC Press LLC, Atlanta, Georgia. 1999

Nakadai, K.; Mizumoto, T., Nakamura K. “Robot-Audition-Based Human-Machine Interface for a Car”, IEEE, 2015.

Ogata, Katsuhiko. “Ingeniería de control moderna”. Ed. Prentice Hall, España 2010.

Ortiz Uribe. “Diccionarios de Metodología de la Investigación Científica”. Segunda Edición. Ed, Limusa, Méx. 2008.

Proakis, John G.; Manolakis, Dimitris G. “Tratamiento Digital de Señales”, Ed. Pearson Educación, Madrid, 2007.

Rabiner, Lawrence.; Schafer, Ronald W. “Digital Processing of Speech Signals”, Ed. Prentice Hall, Englewood Cliffs, New Jersey 1978.

Schmidt, Maik. “Raspberry Pi: A Quick-Start Guide”, Ed. The Pragmatic Programmers, LLC. United States of America 2014.

Smith, Steven W. “The Scientist and Engineer's Guide to Digital Signal Processing”, Ed. California Technical Publishing, United States of America 1999.

Varga, I. Aalburg; S. Andrassy B; Astrov S; Bauer J.G; Beaugeant C; Geißler, C. - Höge H. “ASR in Mobile Phones—An Industrial Approach”, IEEE, 2002.

Vaseghi, Saeed V. “Multimedia Signal Processing”, Ed. John Wiley & Sons Ltd, England 2007.

White, Glen. “Introducción al Análisis de Vibraciones”, Ed. Azima DLI, U.S.A 2010.

W, Weisstein Eric. “CRC Concise Encyclopedia of Mathematics”, Ed. CRC Press LLC, Washington, D.C 1998.

Tesis consultada:

Mehta, M. G. “Speech Recognition System”, 1996.