Diseño de un algoritmo para un sistema de procesamiento de comandos de voz para el control de sillas de ruedas

William Rene Inguilan Ceballos

Universidad Antonio Nariño

Facultad de Ingeniería Mecánica, Electrónica y Biomédica

Pasto, Colombia

2018

Diseño de un algoritmo para un sistema de procesamiento de comandos de voz para el control de sillas de ruedas

William Rene Inguilan Ceballos

Proyecto de grado monográfico presentado como requisito parcial para optar al título de:
Tecnólogo en electromecánica

Director (a):
Ingeniero Luis Enrique Arteaga

Línea de Investigación:
Automatización

Universidad Antonio Nariño
Facultad de Ingeniería Mecánica, Electrónica y Biomédica
Pasto, Colombia
2018

Agradecimientos

En primer lugar agradezco a nuestro creador ya que sin la ayuda de él no es posible alcanzar nuestras metas, él es nuestro guía y una protección, quien nos enfrenta a los retos necesarios para podernos superar. Agradezco a mi familia por el sacrificio y apoyo incondicional, por ser una motivación para seguir adelante y que ellos puedan sentirse orgullosos de nuestros logros. Agradezco a mis maestros por los conocimientos trasferidos, por la motivación y ejemplo emitido, a nuestros compañeros por compartir las vivencias alegres y difíciles de nuestro aprendizaje y gracias a todas las personas que han aportado desde lo más minino para que podamos llegar a este nivel.

Resumen

En este proyecto se lleva a cabo una investigación sobre sistemas para el reconocimiento de comandos de voz, que se adapten a una silla de ruedas, este enfoque se lleva a cabo teniendo en cuenta la población con discapacidad motora, la cual se ve afectada por la sociedad excluyente en cuanto a: brindar espacios de interacción, servicios necesarios, trabajo, etc. El fin de este trabajo radica en la elaboración de un algoritmo que resuma las metodologías optimas, extrayendo las técnicas necesarias para un buen procesamiento de señales acústicas, capaz de filtrar las señales de frecuencia inaudible y otras perjudiciales para el ser humano, donde el rango de utilidad es desde 20Hz a 20kHz, este algoritmo es capaz de reconocer no menos de 12 características fundamentales del espectro de la voz, para poder identificar así los fonemas de las palabras. Entre los métodos que son parte del sistema se encuentra los Modelos ocultos de Markov, que son comparaciones estadísticas.

Palabras clave: comandos de voz, reconocimiento, procesamiento de señales, Markov.

Abstract

In this project an investigation is carried out on systems for the recognition of voice commands, which is adapted to a wheelchair, this approach is carried out taking into account the population with motor disability, which is affected by the Exclusive society in terms of providing services, spaces, work, etc. The purpose of our work lies in the development of an algorithm that summarizes the optimal methodologies, extracting the necessary techniques for a good processing of acoustic signals, capable of filtering inaudible and harmful frequency signals for the human being, where the range of usefulness is from 20Hz to 20kHz, this algorithm is capable of recognizing no less than 12 fundamental characteristics of the voice spectrum, in order to identify the phonemes of the words. Among the methods that are part of the system is the Hidden Markov Models, which are statistical comparisons.

Keywords: voice commands, recognition, signal processing, Markov.

Contenido

Pág.

Lista de figuras

Pág.

Lista de ecuaciones

Lista de tablas

Pág.

Introducción

Hasta hace 4 años según la Organización Mundial de la Salud (OMS, 2014) en el mundo existen más de 1000 millones de personas con discapacidad, alrededor de 15% de la población total, lo cual va en crecimiento debido al envejecimiento de las poblaciones y la proliferación de enfermedades crónicas, este problema se acentúa en los países con menor desarrollo económico, como lo es Colombia lugar donde además las personas con discapacidad se enfrentan a dificultades para acceder a los servicios básicos, como sanidad, educación especializada y demás, necesarios para garantizarles una calidad de vida. Más aun cuando el tipo de discapacidad le impide movilizarse libremente, dependiendo de al menos una persona para realizar sus actividades.

Este problema ha llevado a las investigaciones a enfocarse en la obtención de soluciones que le brinde mayor autonomía a las personas con discapacidad, para que así ellos puedan desenvolverse en muchos ámbitos y sean incluidos por la sociedad a nivel de servicios y trabajo. Esta investigación permite analizar mejor algunos de los estudios realizados en cuanto al reconocimiento de comandos de voz para poder aplicarlos al control de sillas de ruedas, se trata de aportar desde los aspectos técnicos y metodológicos una estructura de algoritmo para el procesamiento de señales, así estos aportes sean en adelante una base para la aplicación de soluciones relacionados con la salud y la sociedad. En el aspecto técnico general esta investigación conlleva a que los futuros diseñadores de sillas de ruedas tengan en cuenta en sus prototipos la implementación de los sistemas de reconocimiento de voz y se facilite la estructuración de los mismos.

Las temáticas que se tratan en este documento se basan en datos teóricos y prácticos de sistemas construidos, en cuanto al tratamiento de las señales acústicas, lo primero a tener en cuanta será digitalizar las señales dentro de una frecuencia manejable de 8000Hz, establecerla dentro de unos parámetros de normalización de 0 a 1, la filtración de señales de ruido que se encuentren por fuera del rango entre 20Hz a 20kHz frecuencia audible para los seres humanos (Casas, Cruz, & Jurado, 2017), obtener datos manejables, por ejemplo “palabras” que es como se le llama a la unidad mínima de reconocimiento y así extraer de ellas no más de 12 características para identificar los comandos emitidos por la voz humana, filtrado de los ruidos inmersos en la frecuencia audible y demás perturbaciones del ambiente. Establecer las características mas apropiadas y más usadas, de este modo dar un orden lógico e ideal para solucionar la problemática sobre el reconocimiento de comandos de voz que sea aplicado para el control de una silla de ruedas.

Este trabajo se basa en una estructura complementada la incluye una entrada de datos, pre-procesamiento, extracción de características, reconocimiento de comandos de voz a través de modelos ocultos de Markov y salida de datos de resultado.

El objeto principal será la consecución de un algoritmo en base a esa estructura definida para cumplir con cada componente de un mismo sistema, representara el fin de este trabajo y concluye como resultado de la investigación.

Nos limitamos al entorno investigativo de la metodológica empleada, pero que dará pautas para la inclusión de nuevos aportes y se pretende publicar de forma abierta para aquellos fabricantes, estudiantes y demás personas interesadas en estos temas.

1. Objetivos

1.1 Objetivo general

El principal objetivo de este trabajo es diseñar un algoritmo basado en una estructura definida que permita el procesamiento de comandos provenientes de señales acústicas de voz humana que puedan ser aplicados al control de una silla de ruedas.

1.2 Objetivos específicos

Elaborar una revisión bibliográfica con contenidos basados en el procesamiento de comandos de voz.

Analizar las técnicas utilizadas para adecuación de las señales acústicas para su interpretación.

Analizar las técnicas utilizadas en el procesamiento de señales acústicas para la extracción de características de voz.

Implementar en el reconocimiento de voz permitiendo reducir errores en los resultados.

Emitir un algoritmo para el reconocimiento de comandos de voz aplicada al manejo de sillas de ruedas.

2. Metodología

Considerando la masiva información referente al procesamiento de comandos de voz y teniendo en cuenta la gran cantidad de datos utilizados para el desarrollo de aplicaciones específicas en estas áreas, se expone a continuación la metodología utilizada en el proceso investigativo, el análisis sobre las técnicas teóricas, descripción y algoritmos de procesamiento de comandos, basada en la metodología de (Martí, 2017).

1. Etapa 1

   Aquí se identifica el problema y demanda sobre el producto de este trabajo, para lo cual se adentra en la cotidianidad de las personas discapacitadas y la sociedad, dirigiéndose a recientes investigaciones.

2. Etapa 2

   Se realiza una recolección de información de distintas fuentes bibliográficas para determinar los múltiples factores que influyen en el procesamiento de comandos de voz. La investigación se desarrolla con un enfoque cuantitativo teniendo en cuenta la realización del algoritmo para el procesamiento de comandos de voz que pueda ser implementado en el control de sillas de ruedas.

3. Etapa 3

   Se establece el contexto actual de implementación de reconocimiento de comandos de voz, encontrados en los recientes estudios que hayan logrado la identificación y aplicación de estos sistemas, logrando identificar la estructura base, métodos y sus resultados.

4. Etapa 4

   Aquí se define una estructura de algoritmo la cual tiene un orden lógico y se basa en las técnicas ideales para el reconocimiento de comandos extraidas de las etapas anteriores.

Tabla 1: Cronograma de desarrollo de las etapas de trabajo

ETAPA	ACTIVIDAD	SEMANAS									
		1	2	3	4	5	6	7	8	9	10
1	Identificación del problema o necesidad	X	X								
2	Investigación, recolección bibliográfica de información específica de aplicación, análisis y estudios.			X	X						
3	Investigación del contexto actual en cuanto a reconocimiento de comandos de voz.					X	X				

4	Emisión de una estructura de algoritmo	X	X

3. Planteamiento del problema

3.1 Discapacidad y tecnología

Esta parte de la investigación demuestra la importancia del enfoque de apoyo a la autonomía para las personas con discapacidad motriz a quienes este tipo de avances en tecnología es lo que les permite alcanzar una mejor calidad de vida y ser incluidos en la sociedad. La inactividad motriz de las personas se debe a múltiples factores entre los que se pude mencionar falencia del sistema nervioso en particular el cerebro y traumas (A. M. Martínez, Alcaraz, J. R. C V, & L, 2018),

Ya que el entorno de nuestra vida es excluyente y se tiene en cuenta otros aspectos que conllevan a la discapacidad que el mismo entorno de aquellos discapacitados, se han direccionado los esfuerzos de apoyo a conflictos como narcotráfico, desnutrición, crecimiento de la población, envejecimiento, accidentes de tránsito, etc.

En un ambiente de prejuicios donde se asocia a un discapacitado con conceptos de invalides, se asume un producto o servicio accesible para un fin colectivo y se tiene en cuenta solo a la mayoría, es decir no se incluye a personas con alguna discapacidad. (Coronel, 2016).

Es ahí donde entran las tecnologías, para reunir esfuerzos por darle un bienestar a estas personas, en especial a aquellas que no cuentan con la forma de movilizarse por sí mismas, dándoles una nueva opción de mejorar, permitiendo avanzar en la inclusión de la sociedad y que en su vida personal dependan menos de otras personas para desarrollar sus actividades.

Actualmente se han implementado mejores y más novedosos sistemas que interactúan entre máquina y ser humano, los que coadyuvan a mejorar diferentes apartes de la medicina, la seguridad y muchos espacios de la vida cotidiana, se ha logrado desarrollar sillas de ruedas que obedecen las órdenes de los comandos de voz de un usuario, mediante un sistema de reconocimiento de voz específico basado en la transformada de Fourier discreta para el estudio (A. M. Martínez, Alcaraz, J. R. Cárdenas Valdez, & López, 2018).

Además, algunos de los científicos e ingenieros están desarrollando diferentes aplicaciones electrónicas con el fin de mejorar la calidad de vida en los seres humanos con incapacidad motriz, logrando una ejecución tecnológica para adecuarla a todo tipo de entornos y permitiendo el control de máquinas por medio de la voz (A. M. Martínez et al., 2018).

Todos estos avances se realizan basados en técnicas enfocadas en muchos aspectos como el reconocimiento de comandos de voz, que es en el que enfatizamos para aplicarlo a una silla de ruedas, discerniendo las técnicas más optimas posibles que permitan un control sin errores.

En nuestro entorno (la ciudad de Pasto, departamento Nariño, república de Colombia) encontramos que se vienen realizando adelantos en la educación tecnológica pero que aún no es suficiente, además los recursos necesarios para aplicaciones avanzadas son de difícil acceso y a precios inalcanzables, por lo que recurrimos a la investigación de sistemas creados que nos permitan dar alternativas a los diferentes retos que se presentan, formas más accesibles para personas de nuestra sociedad.

4. Marco teórico

Esta etapa del trabajo resume el análisis de las bibliografías estudiadas, para dar claridad a la temática sobre el reconocimiento de comandos de voz, enfatizando en los conceptos necesarios que intervienen en el proceso que se le da a las señales de audio, desde su generación que provienen de la voz humana su estructuración y como se convierte finalmente den una acción mecánica.

4.1 Fisiología de la voz

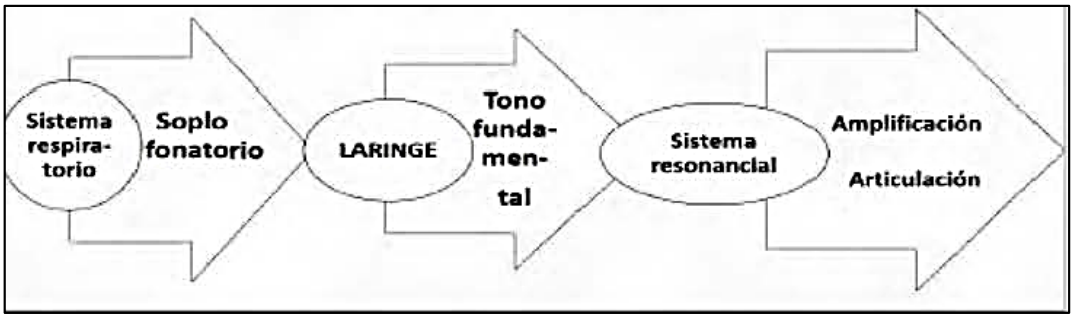
La voz humana es controlada desde una sección del cerebro humano conocida como "Broca", se ubica en el hemisferio izquierdo de su corteza y hace parte del sistema nervioso central (Pérez, Poceros, & José, 2013), de ahí se realiza el envío de estímulos hacia la laringe, donde se encuentra con la glotis, se generan los sonidos y pasa el aire que viene de los pulmones haciendo vibrar las cuerdas vocales o membranas. Estos movimientos producen una resonancia alrededor de la nariz y la faringe.

El aparato de fonación puede ser controlado conscientemente por el ser humano en cuanto trata de comunicarse, la variación de la intensidad de los sonidos depende de la fuerza de la espiración. En el hombre las cuerdas vocales son algo más largas y gruesas que en la mujer y el niño, por lo que produce sonidos más graves.

En la Figura 1 podemos observar el instrumento fonador como una secuencia de sus partes y las transiciones que surgen, también podemos encontrar en la Figura 2 la fisiología que contiene el sistema fonador del ser humano. A continuación la definición de algunos conceptos básicos:

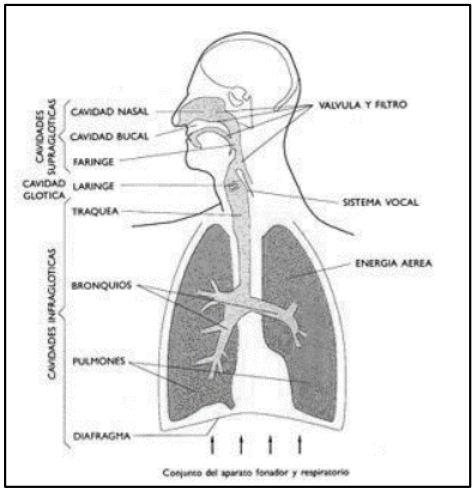
- **Aparato respiratorio**, funciona a través de dos movimientos la inspiración y la espiración, es en este último que logra expulsar el aire para producir el sonido, circulando el aire por la nariz, tráquea, pulmones y diafragma.
- **Aparato fonador**, el aire llega a la laringe hace vibrar las cuerdas, es ahí donde se produce el tono fundamental.
- **Aparato resonador**: compuesto por cavidades nasales, boca y faringe, le imprime de brillo y redondez a la voz, el sonido adquiere sus características dependiendo de forma y posición en cada cavidad, estas son las que diferencian a unos de otros, además está la frecuencia de estas vibraciones.

Figura 1: Componentes del aparato fonador



Tomado de (Angiono, Thompson, Lucini, Serra, & Serra, 2017)

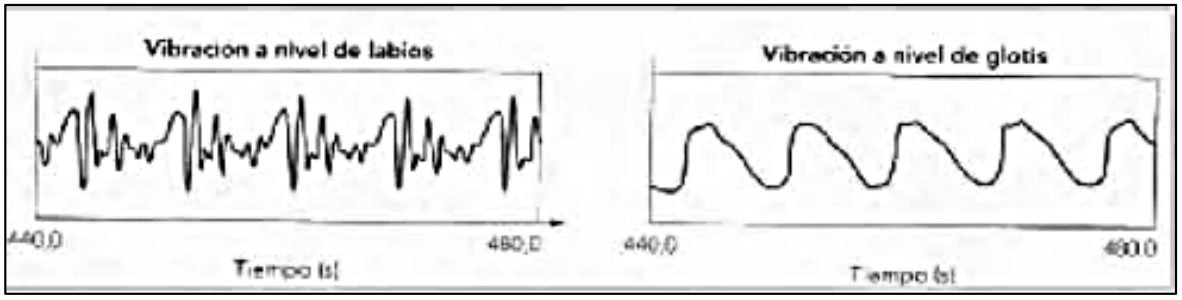
Figura 2: Aparato fonador



Tomado de (Pérez et al., 2013)

Se han realizado múltiples estudios para identificar las diferencias entre las ondas de sonido emitidas, incluso entre aparato fonador y resonador, su observación permite el análisis de comportamiento de la frecuencia, brillo, etc. Se ha podido diferenciar en los parámetros analizados las vocales y consonantes; también se encuentra variedad en las diversas partes del sistema fonador entre hombres, mujeres y niños, los estudios realizados permiten el manejo de parámetros en respiración, resonancia, intensidad, altura vocal, extensión vocal, etc. (Angiono et al., 2017). En la figura 3 se muestra una comparación de las vibraciones en diferentes partes del aparato fonador, lo cual permite establecer cuál es la función que desempeña cada una y el resultado esperado.

Figura 3: Vibraciones a nivel de los labios y a nivel de glotis.



Tomado de (Angiono et al., 2017)

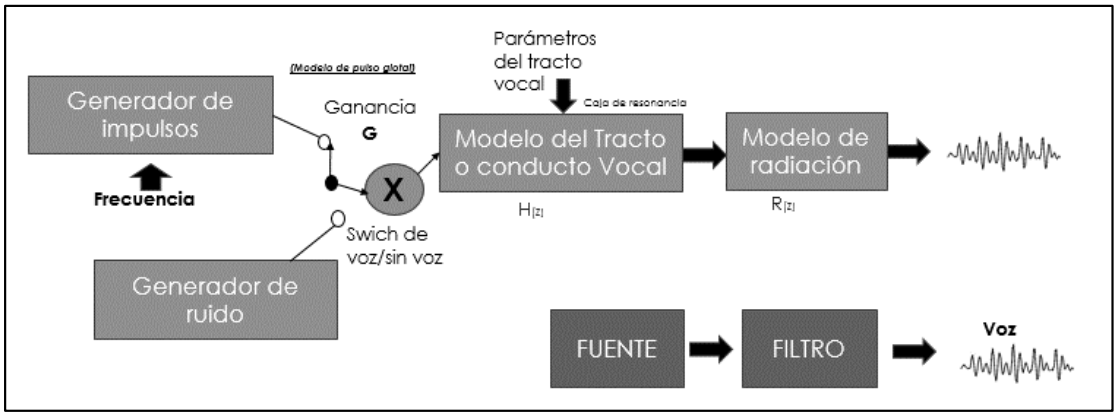
4.2 Clasificación de los Sonidos

La RAE (Real Academia Española) define el sonido como: “sensación producida por el órgano del oído” motivado por el movimiento vibratorio de los cuerpos, trasmitido por un medio elástico, como el aire” (Vela, 2017).

Las principales unidades fundamentales del sonido generado por los seres humanos, son las vocales y consonantes, las primeras son producidas cuando el tracto vocal se excita por pulsos de aire emitidos por las cuerdas vocales, su vibración es periódica a diferencia de las consonantes donde las cuerdas vocales están relajadas, Basado en (Gómez, Simancas, Acosta, Meléndez, & Vélez, 2016).

Podemos sintetizar el proceso para generación de la voz de forma artificial en un esquema de funciones como el de la figura 4, donde se presenta un generador de pulsos necesarios, como lo serian los pulmones, un swich de voz como la .

Figura 4: Esquema de la generación de la voz



Tomado de (Gómez et al., 2016)

4.3 Oído humano

Los seres humanos tenemos la capacidad de escuchar gran cantidad de sonidos al mismo tiempo, la mayor parte de estos son ruidos e interferencias que no son de nuestro interés a la hora de comunicarnos, es por esta razón que el oído también cuenta con formas de filtración que nos permite escuchar lo que en realidad necesitamos. El sonido se considera como vibraciones de partículas y sistemas materiales con masa y elasticidad, las cuales atraviesan los oídos para ser transformados en un tipo de energía que se pueda transferir hacia el cerebro y así ser interpretada. Las vibraciones audibles para el hombre se encuentran en el rango de frecuencias de 16Hz y 20kHz, por debajo de 16Hz (infrasonidos) y por encima de 20kHz (ultrasonidos), ya no se pueden captar por el oído humano, la mayor parte de los sonidos que el ser humano capta en su vida diaria se encuentran entre los 500 y 8000Hz. Otro de los parámetros que define lo que se puede percibir es la intensidad de sonido, donde podemos situar un rango de 0 a 120Db, los valores cercanos a los 120Db empiezan a generar molestias o dolor en los oídos (Angiono et al., 2017).

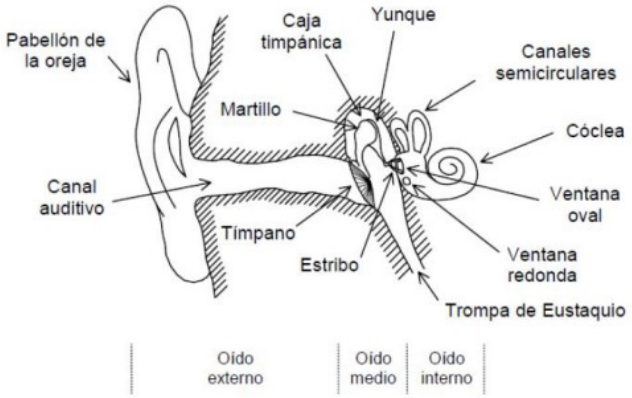
El oído es un órgano sensorial que trabaja como transductor, capaz de convertir las señales sonoras en señales eléctricas, estas señales son enviadas al cerebro, el cual procesa, interpreta y almacena dicha información" (Guamán, 2018), en la tabla 2 encontramos algunas partes y sub partes del oído humano y la figura 5 muestra la fisiología de las mismas:

Tabla 2: Partes del oído humano

oído externo	pabellón	
	conducto auditivo externo	
oído medio	membrana timpánica	
	ventanas oval y redonda	
	trompa de Eustaquio	
	cadena de huesillos	martillo
		yunque
		estribo
oído interno	vestíbulo	sáculo
		utrículo
	canales semicirculares	
	caracol	rampa vestibular
		rampa coclear
		rampa timpánica

Tomado de (Guamán, 2018)

Figura 5: Componentes del aparato fona torio



Tomado de (Guamán, 2018)

4.4 Sistemas para la adquisición de señales

Desde tiempos antiguos diferentes autores han identificado la necesidad de representar parámetros físicos como por ejemplo temperatura, humedad, presión, caudal, etc. de forma cuantificada con diferentes métodos descriptivos. Los avances tecnológicos permiten que muchos de los sistemas que necesitan entradas de señales analógicas puedan transformarlas como una base de información para el posterior procesamiento y análisis, el resultado de esto será una toma de decisiones que puede conllevar a unas acciones físicas. Las señales analógicas se encuentran en todo nuestro entorno físico y se han representado de diferentes formas y bajo diferentes métodos para obtener una descripción lo más parecida a ellas.

La rama de la instrumentación trata de percibir esa realidad utilizando diferentes métodos para la extracción de información del entorno mediante la identificación y cuantificación de las características físicas suficientes para describir diferentes estados que se presentan cotidianamente (Ochoa, 2018). En este caso las señales analógicas a identificar son señales de acústicas y en específico aquellas generadas por la voz humana.

4.4.1 Características de las ondas de sonido

Según (Lopez, 2018) la onda de sonido es un tipo de onda mecánica de desplazamiento, la que se propaga de forma longitudinal transportando energía, a través de materiales elásticos o viscosos ya sea solido liquido o gaseoso, las ondas de sonido

se propagan a través del aire y hacen vibrar sus partículas ocasionando cambios de presión y densidad en dirección de las mismas, la medida adoptada para cuantificar estas ondas es la presión sonora, sus unidades son newton por metro cuadrado (N/m2), existe una relación con la presión atmosférica representada como:

$$\mathrm{p(t)} = \mathrm{P(t)} - P(atm) \qquad \text{Ecuación 1}$$

Donde t es el tiempo, p (t) la presión sonora en el instante de tempo t, P (t) la presión momentánea del aire en el instante de tiempo t y P (atm) la presión atmosférica.

Otra característica de las ondas es la potencia la cual se entiende como la cantidad de energía por unidad de tiempo, la potencia se mide en un punto fijo (intensidad sonora o densidad de potencia), sobre un área infinitesimal de tamaño dA

$$\mathrm{I} = \frac{dP}{dA} \qquad \text{Ecuación 2}$$

Donde dP es la potencia acústica en el área dA detectada por el sensor, estará perpendicular a la dirección de la onda sonora, la intensidad sonora se mide en watts por metro cuadrado (W/m^2).

4.5 Estructura para el reconocimiento de comandos de voz

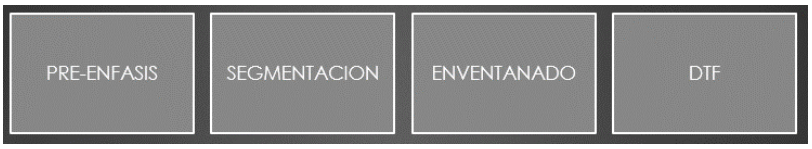
4.5.1 Pre-procesamiento

Esta es la primera etapa general que conforma la estructura del sistema de reconocimiento de voz en la que se dice que la señal de salida de un sistema externo se debe procesar de una forma adecuada para la siguiente etapa de la operación, La señal puede ser por ejemplo, amplificada; podría contener interferencias que eliminar; ser análoga y requerir su digitalización o viceversa; requerir un cambio de voltaje etc.

Los elementos que conectan un sistema electrónico con su entorno muchas veces no están preparados para ser conectados con el núcleo del sistema, las etapas de acondicionamiento de señal, hacen compatibles dichas conexiones. En este sentido las muestras tomadas necesitan de un pre-procesamiento de sus señales, con el fin de adecuarse al sistema y así permitir interpretar y procesar mejor los datos.

La figura 6 nos muestra las partes internas que se presentan normalmente en el pre-procesamiento de las que se profundiza a continuación.

Figura 6: Estructura interna de pre procesamiento



Basado en (Gordillo, 2018)

4.5.2 Pre énfasis

Los acondicionadores de señal consisten en circuitos que transforman los parámetros eléctricos de salida de los transductores en una señal eléctrica (corriente, voltaje o frecuencia) la cual puede medirse de forma factible, este proceso previo al procesamiento de extracción de la información deseada, parte de la instrumentación electrónica, transductores y acondicionadores de señal (Miguel & Bolad, 2015, pag 6-7).

Razones para el acondicionamiento de señales

- Cuando la señal eléctrica medida no está dada en magnitud o intensidad, por lo cual es conveniente un circuito que realiza esta conversión Convertir una señal en un tipo de señal adecuado.
- Protección para evitar el daño al siguiente elemento.
- Obtención del nivel adecuado de la señal, en este caso el acondicionamiento implica: aumentar la señal a niveles superiores al ruido electrónico, y filtrar señales para eliminar el ruido por interferencia.
- Eliminar la baja amplitud que por lo general representa espacios en silencio.
- Manipulación de la señal. Por ejemplo, convertir una variable en una función lineal, mediante circuitos específicos o programas de cálculo adecuados.

Se utiliza un filtro de alto orden para enfatizar previamente la muestra de voz (Kavitha, Nachammai, Ranjani, & Shifali, 2014).

$$Y(d) = X(d) - C * X(d-1) \qquad \text{Ecuación 3}$$

Donde Y (d) es la señal de salida, y X (d) es la señal de entrada C es una constante con un valor entre 0.9 y 1.

4.5.3 Segmentación

En la segmentación se divide el vector de la señal acústica en tramos de entre 20 a 40 ms para realizar el análisis de cada tramo de forma individual, lo más usual es usar 25 ms, la frecuencia para este tipo de señales es comúnmente 8 kHz o mayor. Se puede distribuir estos cuadros de análisis de modo que se realice una superposición a fin de no perder información entre los bordes de inicio y final de cada cuadro. Uno de estos métodos se le denomina ventana de Hamming donde la superposición vuelve a colocar las características de la señal en las derivadas el tamaño del cuadro esta generalmente en las potencias de dos.

4.5.4 Enventanado

Proceso en el que se integran los cuadros para evitar distorsiones y discontinuidades en las partes inicial y final de cada cuadro. Cuando se usa la transformada de Fourier en estas muestras de señales acústicas se debe realizar las superposiciones a fin de que las transiciones de señal sean lo más suaves posible y facilite el inicio y termino de cada uno, comúnmente se utiliza para el reconocimiento de comandos la ventana de Hamming H(d).

$$H(d) = 0.54 - 0.46\ cos(\frac{2\pi d}{D-1}) \qquad \text{Ecuación 4}$$

$$0 \leq d \leq D-1$$

La señal procesada se obtiene por la relación de:

$$Y(d) = X(d) * H(d) \qquad \text{Ecuación 5}$$

Aquí D = número de muestras en cada cuadro, Y (d) = señal de salida, X (d) = señal de entrada

4.5.5 Extracción característica

Este proceso consiste en derivar las muestras de señal para encontrar unos coeficientes los cuales describen la señal de entrada de otra forma, existen diferentes técnicas de extracción de características de una señal, como lo son los coeficientes de cepstral de frecuencias Mel (MFCC), los coeficientes de cepstral de factor humano (HFCC), coeficientes de cepstral predictivos lineales (Kavitha et al., 2014). Para el caso de reconocimiento de comandos se trabajara con los MFCC debido a que tienen menor complejidad y aun así ofrecen resultados óptimos. De la extracción de estas características los valores correspondientes la descripción de la onda acústica se convierte en coeficientes cepstrales de frecuencia de Mel(A. M. Martínez et al., 2018). El proceso se realiza mediante tres pasos comunes: Transformada de Fourier, Banco de filtros de Mel y Transformada de cosenos discreta de Fourier.

“Con esta técnica se reduce un gran número de datos de la señal grabada, describiendo así las propiedades más importantes de la señal, tales como: tiempo para calcularlas, espacio para almacenarlas, facilidad de implementación” (Pérez et al., 2013)

- **Transformada de Fourier**

En 1807, Jean Baptiste Joseph Fourier físico-matemático francés, demostró que una función podría ser desarrollada en términos de series trigonométricas, y que podían obtener por integración, fórmulas para los coeficientes de Fourier, siendo al día de hoy utilizada en el procesamiento y análisis de señales, con resultados satisfactorios en el caso en que las señales son suficientemente regulares y periódicas (Maldonado, 2018).

Es así como la transformada de Fourier pasa del dominio del tiempo al dominio de la frecuencia en cada tramo haciendo lo siguiente. Cuando tomamos N puntos de una señal para analizar, estamos implícitamente multiplicando por una ventana rectangular.

$$WR(n) =$$

$X(k) = \sum_{n=1}^{N} x(n)h(n)e\hat{}j2\pi kn / N$, 1≤ k ≤ K Ecuación 6

X(k) = valores espectrales, H(n) ventana rectangular, h(n) = es una ventana de análisis de muestra N longitud. K = Es la longitud de la Trasformada Discreta de Fourier.

▪ Banco de filtros de Mel

Se toman grupos que contengan un periodo y se suman para obtener una idea de energía, presente en cada región de las frecuencias(Ceballos, Serna-morales, Prieto, Gómez, & Redarce, 2011).

Para reconocimiento voz solo son necesarios los primeros 12-13 coeficientes denominados coeficientes cepstrales en las frecuencias de Mel (Lyons).

Imitando la forma en que los humanos escuchan, la escala de frecuencia Mel, tiene un esparcimiento lineal de frecuencia por debajo de los 1000 Hz, las señales de voz tienen más energía en las frecuencias más bajas, las siguientes formulas son para calcular los mels de una frecuencia dada en Hz.

$mel\ (f) = 2595.^{10} \log\ (1 + \frac{j}{700})$ Ecuación 7

Para cada tono con una frecuencia actual *f* Hz, un tono subjetivo se mide en la escala de Mel. El pitch de un tono de 1 kHz, 40 dB por encima de la audiencia perceptual se conoce entonces como 1000 mels. En este caso sería:

$mel\ (f) = 2595.^{10} \log\ (1 + \frac{j}{700}) = 181.312.111{,}042623$ Ecuación 8

▪ Transformada de cosenos discreta de Fourier

Se obtienen los coeficientes evaluados que representan la señal de voz normalizada. N, es el número de filtros triangulares y mj son los coeficientes a la salida del banco de filtros.

$Ci = \sqrt{\frac{2}{N}} * \sum_{j=1}^{N} mj * \cos(\frac{\pi}{N} * (j - 0.5))$ Ecuación 9

4.5.6 Reconocimiento de voz

Al proceso de convertir una señal acústica a una secuencia de palabras en forma de texto por medio de un dispositivo de control se le llama reconocimiento de voz. Según el artículo de Sistema audiovisual para reconocimiento de comandos (Ceballos et al., 2011) existen diversas formas de aplicar el reconocimiento de voz, algunos de los más usados son los HMM (Modelos Ocultos de Markov), reconocimiento usando solo audio y reconocimiento del habla basado solo en fonemas.

Figura 7¡Error! No se encuentra el origen de la referencia.: Estructura interna de procesamiento

Basado en (Gordillo, 2018)

- **Tiempo dinámico de pandeo (DTW)**

Técnica DTW es utilizada para determinar y comparar las distancias entre una y otra curva de señales, el resultado de esta presenta un modelado no bien definido, razón por la que en estos momentos no es tan empleada, requiere de segmentos básicos de palabras, usado para reconocimientos aislados, el procesamiento aumenta en cuanto aumenta la base de datos de modelos de entrenamiento, convirtiéndose en una desventaja en el momento de aumentar la unidad de reconocimiento (Ureña, 2011).

- **Modelos ocultos de Markov (HMMs)**

Es un método de reconocimiento estadístico de señales el cual es empleado comúnmente. Los HMMs son modelos con secuencias estadísticas en los que se procesan señales estacionarias, estos representan una secuencia de tiempo y variación de la magnitud de voz, un

HMM describe una palabra como una transición estados en un solo sentido, para más de 20 estados hay un crecimiento logarítmico de la probabilidad (Ceballos et al., 2011).

Estos fonemas recolectados estarán acompañados de su correspondiente representación como serie de palabras, cuanta más información de modelo se tenga el reconocimiento será más exacto, los modelos de lenguaje tendrán información sobre las palabras y las posibles combinaciones que se deben realizar, para esto se requiere de gran cantidad de datos, por lo que se opta por usar aproximaciones como las basadas en N-Gramas asignando probabilidad a posibles próximas palabras, así estos se podrían usar para probabilidades de frases enteras, estos términos provienen de los modelos de Markov. A esto se relacionan las características de señal del hablante, como pueden ser estilo, tono y ritmo del habla, las características son únicas para cada persona ya que el tracto vocal y fisiología son diferentes al comparar sus variaciones de frecuencia, se debe filtrar las señales de ruido ya que son principal causa que impide la identificación de las palabras (Gil Vásquez, Castillo Ossa, & Flórez Hurtado, 2017).

- **Reconocimiento de voz usando solo audio**

Un paquete de herramientas utilizadas en el reconocimiento de voz capaces de manipular diferentes formatos de archivos de audio e incluso algoritmos para la extracción de características acústicas como los índices de Mel. Aspectos que se deben considerar en el reconocimiento de voz HTK: sintaxis de palabras teniendo en cuenta el orden de las mismas, definir un diccionario de palabras con representación en cadenas de Markov (fonemas), obtener las señales de audio con etiquetas de fonemas, definición y extracción de características correspondientes a los 12 índices de Mel, la energía de las señal y las primeras dos derivadas temporales de los índices, el análisis se hace en ventanas de 20ms con traslape del 50%, debido a que las características dependen de la amplitud de la señal se debe normalizar con respecto a la máxima amplitud antes de calcular los índices de Mel; creación y entrenamiento de los modelos, donde se crea primero el prototipo de cadena de Markov, tres estados con propagación hacia adelante para reducir los cálculos y numero de parámetros.

- **Reconocimiento del habla basado en fonemas**

Una base de datos debe ser acondicionada, para esto se segmenta en unidades básicas, es la aproximación más natural utilizada para el reconocimiento de señales de audio ya que es la forma en la que el ser humano reconoce la señales de voz, uniendo sonido tras sonido, filtrando los demás sonidos que se encuentran en el ambiente y separando palabras unas que en voz natural se unen. Otra de las ventajas es que las palabras no necesitan estar dentro de un diccionario, ya que simplemente con el modelo se hace el reconocimiento, pero se debe contener una amplia base de datos de entrenamiento en cuanto a modelos de fonemas.

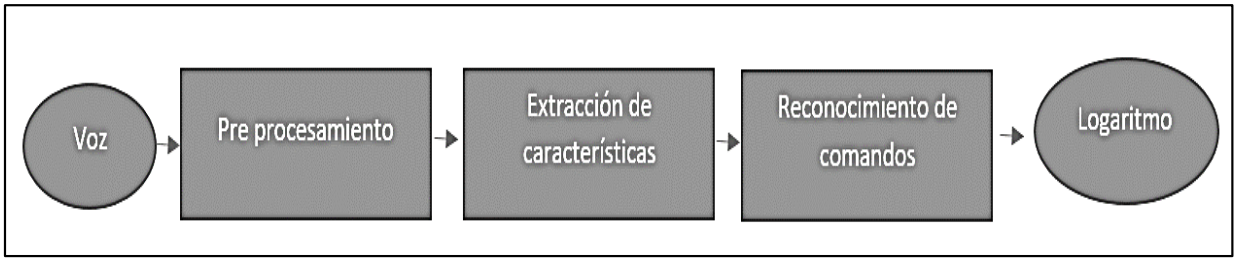
5. Resultados

5.1 Estructura de algoritmo para el reconocimiento de voz para silla de ruedas

Esta investigación ha llevado a este punto en el que se define una estructura de algoritmo para el reconocimiento de comandos de voz más común y con los elementos necesarios sin dejar de ser efectivos para ser aplicados en una silla de ruedas.

Lo anterior se basa en la implementación de un algoritmo con algunas tareas específicas como la realización del prototipo (el cual en este caso no desarrollará), se da en el entrenamiento de code-book u oculto de Markov de cada palabra y reconocimiento de palabras aisladas (O. I. H. Martínez, Flórez, & Medina, 2016).

Figura 8: Estructura de un sistema de reconocimiento de voz



Basado en (Gordillo, 2018)

5.1.1 Pre procesamiento

- **Preénfasis**

Se utiliza un filtro para bajas frecuencias por debajo de los 20Hz y uno de alta por encima de los 20kHz previo a la muestra de voz.

$$Y(d) = X(d) - C * X(d-1)$$

Donde X (d) es la señal de entrada, Y (d) es la señal de salida y C es una constante con un valor entre 0.9-1.

Luego se filtran las señales inferiores a 20Hz y superiores a 20kHz.

- **Segmentación**

En la segmentación para nuestro caso se aplica convenientemente cuadros de 25ms y de 8kHz, como se especifica en 3.3.2. Permitiendo establecer el tiempo y la amplitud justos para el procesamiento de las señales.

- **Enventanado**

Para este caso se aplica las ventanas de Hamming las cuales llevaran a cabo un realce de las características de los segmentos súper posicionados para evitar la pérdida de datos en los bordes de cada 25ms como se ha escogido estableció.

5.1.2 Extracción de características

Se ha escogido para la extracción de características los modelos HMMs, ya que nos ofrece buenos resultados de comparación y es de menos complejidad.

"Para lograr dicho resultado, la señal de voz ingresa a un módulo de procesamiento de señales en el que se extraen los vectores de características sobresalientes que son enviados posteriormente al decodificador" (Gil Vásquez et al., 2017).

Los HMMs y las ANNs presentan habilidades complementarias para el reconocimiento automático del habla, lo que se refleja en las prestaciones superiores de los sistemas híbridos (Ureña, 2011).

5.1.3 Reconocimiento de voz

Para solucionar el reconocimiento de voz en palabras aisladas hemos considerado la implementación de modelos ocultos de Markov debido a la baja complejidad ideal para aplicar a esta unidad de identificación (palabras), además de esta técnica valoramos el uso de descripciones estadísticas que mejora la capacidad del sistema y reduce los errores en comparación con otros métodos como lo es DTW.

5.2 Elaboración de un algoritmo para el reconocimiento de comandos de voz

5.2.1 Redacción

Entrada: Señal de voz

Salidas: Comandos de voz

1. Captura de señal de voz. X(t)
2. Digitalización X(n)
   2.1. Muestreo: es discretica la señal con un periodo t
   2.2. Cuantificación: dar un valor n a la amplitud de la señal muestreada
   2.3. Codificación: utilizar un numero de bits para representar cada uno de los valores cuantificados
3. Filtros de pre-énfasis (filtro pasa bajas)(1-αZ^-1) donde 0.7<α<0.9
4. Segmentación: dividir en cuadros de 25ms a 8kHz con 10ms de superposición.
5. Enventanado: aplicar ventana de Hamming para realce de la parte central del fonema
   x[n]: señal de voz

   w[n]: ventana de análisis

N: Tamaño de ventana

M: Desplazamiento

W[M-n] w [2M-n] w[3M-n]

6. Extracción de características MFCC
   6.1. Aplicar Transformada de Fourier a cada segmento
   6.2. Pasar cada segmento por el banco de filtros de Mel
   6.3. Aplicar logaritmos de la energía:
   - La energía de la señal varia en el tiempo:
     Fonemas sordos menos energía que fonemas sonoros

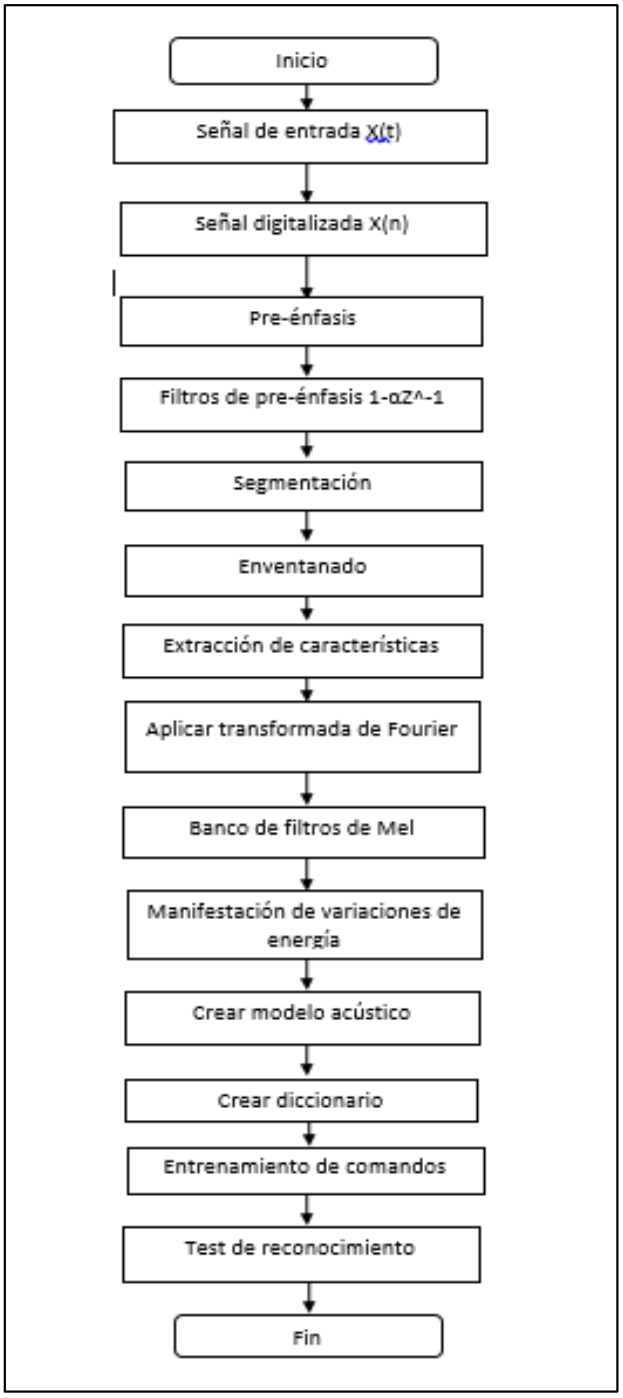
     Consonantes sonoras menor energía que vocales
   - La Energía de tiempo corto pone de manifiesto estas variaciones:
     $En\ \sum_{m-\infty}^{\infty}(x(m)w(n-m)$^2 $En\ \sum_{m-\infty}^{\infty} x^{2(m)}.h(n-m)$
   - Magnitud promedio de tiempo corto:
     1Es una medida alternativa a la energía de tiempo corto

     Es menos sensible a la amplitud de las muestras

     Mn $\sum_{m-\infty}^{\infty}|x(m)|w(n-m)$
7. Crear modelo acústico basado en Módulos ocultos de Markov HMM
8. Crear diccionario de los principales comandos de voz
   1. Adelante = [a ðe 'laN te]
   2. Atrás = ['a tras]
   3. Derecha = [de 're tʃa]
   4. Izquierda = [iθ 'kjer ða]
   5. Pare = ['pa re]
9. Enteramiento de comandos de voz.
10. Test de reconocimiento.

5.3 Diagrama de flujo

Figura 9: Diagrama de flujo


Inicio
Señal de entrada X(t)
Señal digitalizada X(n)
Pre-énfasis
Filtros de pre-énfasis 1-αZ^-1
Segmentación
Enventanado
Extracción de características
Aplicar transformada de Fourier
Banco de filtros de Mel
Manifestación de variaciones de energía
Crear modelo acústico
Crear diccionario
Entrenamiento de comandos
Test de reconocimiento
Fin

6. Conclusiones y recomendaciones

6.1 Conclusiones

La investigación basada en comandos de voz, permitió interiorizar diferentes tipos de diseño de algoritmos con el fin de que se adapte a una silla de ruedas, logrando conocer el sistema de procesamiento de señales y las implicaciones físicas, lógicas y matemáticas que conllevan a un perfecto funcionamiento de un algoritmo.

Implementar el desarrollo de procesamiento de señales con el fin de coadyuvar a la humanidad en la solución de problemas que comprometen la movilidad y motricidad de discapacitados, logrando con esto que se integren a la sociedad y desarrollen funciones acordes a sus capacidades.

Teniendo en cuenta que el aprendizaje a futuro será retroalimentarse y crear nuevas formas de procesar la voz, que coadyuven mediante métodos, equipos, sensores y guías con el fin de que la medicina siga un cauce de investigación electrónico que ayude a las personas con discapacidad a mejorar su calidad de vida, y que la paraplejia no sea un obstáculo, ya que por medio de las creaciones tecnológicas grandes personalidades a nivel mundial han demostrado que la tecnología es un aliado de un futuro sin obstrucción alguna.

Se ha demostrado a base de investigación que la tecnología basada en procesamiento de señales, nos conlleva al desarrollo de un mundo futurista, que utiliza la inteligencia su propio cuerpo, sentidos y funciones, para llevarlos a su propia articulación y funcionamiento artificial pro ayuda de la humanidad y de todos aquellos con algún de tipo de discapacidad.

6.2 Recomendaciones

Se presentan como una serie de aspectos que se podrían realizar en un futuro para emprender investigaciones similares o fortalecer la investigación realiza.
Continuar con la investigación en procesamiento de señales de voz, con el fin de que se innoven en nuevas prácticas tecnologías que permitan a futuro desaparecer todo estigma y obstáculo para personas discapacitadas.

Bibliografía

Angiono, V., Thompson, M. A. M., Lucini, M. B., Serra, M., & Serra, S. (2017). *Fonoaudiologia, bases de la comunicación humana*. *Journal of Visual Languages & Computing* (Vol. 1).

Casas, A., Cruz, O., & Jurado, J. U. (2017). ¿ Te gustaría grabar tu voz u otros sonidos ?, aquí te damos una idea, *18*, 0–13. Retrieved from http://www.revista.unam.mx/rdu-demo/wp-content/uploads/v18_n8_a2-_Casas-et-al..pdf

Ceballos, A., Serna-morales, A. F., Prieto, F., Gómez, J. B., & Redarce, T. (2011). Sistema audiovisual para reconocimiento de comandos Audiovisual system for recognition of commands, *19*, 278–291.

Coronel, J. J. I. (Universidade de S. P. (Brasil)). (2016). Artículos científicos sobre turismo para personas con discapacidad en revistas Iberoamericanas de turismo . Una propuesta de categorización, *14*, 41–58.

Gil Vásquez, L. J., Castillo Ossa, L. F., & Flórez Hurtado, R. D. (2017). Reconocimiento de comandos de voz en español orientado al control de una silla de ruedas. *Revista UIS Ingenierías*, *15*(2), 35–48. https://doi.org/10.18273/revuin.v15n2-2016003

Gómez, J., Simancas, J., Acosta, M., Meléndez, F., & Vélez, J. (2016). Algoritmo de recocimiento de comandos voz basado en técnicas no-lineales. *Espacios*, *38*(17), 18. Retrieved from http://repositorio.cuc.edu.co/xmlui/handle/11323/904

Gordillo, C. D. A. (2018). *Realce e Reconhecimento de Voz Contínua em Ambientes Adversos*.

Guamán, M. S. C. (2018). Medición del umbral de audición en bajas frecuencias e infrasonido, 95. Retrieved from http://dspace.udla.edu.ec/bitstream/33000/10205/1/UDLA-EC-TISA-2018-21.pdf

Kavitha, R., Nachammai, N., Ranjani, R., & Shifali, J. (2014). Speech Based Voice Recognition System for Natural Language Processing, *5*(4), 5301–5305. Retrieved from https://pdfs.semanticscholar.org/14f6/6f7aaebb56f9b7eb7a008d84afce6708bc12.pdf

Lopez, M. A. (2018). Análisis experimental de la factibilidad del uso de compresión con pérdida en la clasificación automática de sonidos biológicos, (June). Retrieved from https://www.researchgate.net/profile/Manuel_Aguilera4/publication/327447358_Analisis_experimental_de_la_factibilidad_del_uso_de_compresion_con_perdida_en_la_clasificacion_automatica_de_sonidos_biologicos/links/5b900b5945851540d1cc2369/Analisis-

experimenta

Maldonado, C. B. G. (2018). *Desarrollo de algoritmos eficientes para identificación de usuarios en accesos informaticos*. UNIVERSIDAD COMPLUTENSE DE MADRID FACULTAD. Retrieved from https://eprints.ucm.es/46037/1/T39510.pdf

Martí, J. (2017). La investigación - Acción participativa estructura y fases, 37–41. https://doi.org/10.1577/1548-8659(2001)130

Martínez, A. M., Alcaraz, G. E. V., J. R. Cárdenas Valdez, & López, C. E. V. (2018). Manejo de Silla de Ruedas Eléctrica por Comandos de Voz Personalizado, *6*. Retrieved from http://fcqi.tij.uabc.mx/usuarios/revistaaristas/numeros/N12/articulos/150-155.pdf

Martínez, O. I. H., Flórez, H. Y. E., & Medina, D. F. M. (2016). Prototipo de silla de ruedas comandada por vozempleando HMM en un ambiente contolado, (July). https://doi.org/10.19053/1900771X.5121

Miguel, M. G., & Bolad, E. M. (2015). *Instrumentacion Electronica: Transductores y acondicionadores de señal*. Retrieved from http://www.editorial.unican.es/libro/instrumentacion-electronica-transductores-y-acondicionadores-de-senal

Ochoa, B. joosely L. (2018). Analisis de consumo monofasico. Retrieved from https://addi.ehu.es/bitstream/handle/10810/29158/TFG_BerthaLorenzoOchoa.pdf?sequence=1&isAllowed=y

OMS. (2014). Proyecto de acción mundial de la OMS sobre discapacidad 2014-2021: Mejor salud para todas las personas con discapacidad, 27. https://doi.org/EB134/2014/REC/2

Pérez, E., Poceros, F., & Jose, V. (2013). Sistema de Seguridad Por Reconocimiento de Voz, 74. Retrieved from http://tesis.bnct.ipn.mx/dspace/bitstream/123456789/12309/1/Sistema de Seguridad por Reconocimiento de Voz (Tesis de Ingenieria ESIME).pdf

Ureña, R. S. (2011). *Maquinas de vectores soporte para reconocimiento robusto de habla*. Retrieved from https://e-archivo.uc3m.es/bitstream/handle/10016/12577/Tesis_Ruben_Solera_Urena.pdf

Vela, L. A. V. (2017). Reconocedor de palabras aisladas medidante la voz, 92.