RECONOCIMIENTO DE COMANDOS DE VOZ EN ESPAÑOL ORIENTADO AL CONTROL DE UNA SILLA DE RUEDAS

LILY JHOHANA GIL VÁSQUEZ

UNIVERSIDAD AUTÓNOMA DE MANIZALES
MAESTRIA EN MECATRONICA Y CONTROL
MANIZALES
(I COHORTE)
2015

RECONOCIMIENTO DE COMANDOS DE VOZ EN ESPAÑOL ORIENTADO AL CONTROL DE UNA SILLA DE RUEDAS

LILY JHOHANA GIL VÁSQUEZ

Informe final del proyecto de grado para optar al título de Magister en Mecatrónica y Control

Director
M.Sc. Rubén Darío Flórez Hurtado
Coordinador Departamento Electrónica y Automatización
Universidad Autónoma de Manizales

Co-director
Ph.D. Luis Fernando Castillo Ossa
Investigador
Universidad Autónoma de Manizales
Universidad de Caldas

UNIVERSIDAD AUTÓNOMA DE MANIZALES
MAESTRIA EN MECATRONICA Y CONTROL
MANIZALES
2015

DEDICATORIA

Este trabajo se lo dedico primero que todo a Dios por no permitirme abandonar esta meta en los momentos difíciles, a mi madre y a mi padre por su apoyo incondicional, el ánimo dado en los momentos que más lo necesitaba y el ejemplo de vida lleno de buenos valores, persistencia, honestidad, liderazgo y amor familiar.

AGRADECIMIENTOS

Agradezco a mis directores, Rubén Darío Flórez y Luis Fernando Castillo, por sus asesorías, colaboración y apoyo incondicional durante todo el proceso. A los docentes del grupo de investigación en Automática y del grupo de investigación en Ingeniería del Software de la Universidad Autónoma de Manizales por las asesorías recibidas en las diferentes etapas del proyecto. De igual manera agradezco a todas las personas que se ofrecieron como voluntarios en la ciudad de Manizales para realizar las diferentes pruebas de reconocimiento del sistema.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

ABREVIATURAS

Notación	Significado
ADC	Conversión análoga a digital
HMM	Modelos Ocultos de Markov
DSP	Procesamiento Digital de Señales
PCM	Modulación por codificación de pulsos.
MFCC	Coeficientes Cepstrales en las frecuencias de Mel
DFT	Transformada discreta de Fourier
FFT	Transformada Rápida de Fourier
PDF	Función de Densidad de Probabilidad
SAPI	Interfaz de Programación de Aplicaciones de Voz
API	Interfaz de Programación de Aplicaciones
DDI	Interfaz de Controlador de dispositivo
HTK	Hidden Markov Model Toolkit

INTRODUCCIÓN

En el presente proyecto se desarrolla una aplicación computacional que reconoce comandos de voz en español para un vocabulario cerrado e independiente del hablante con una gramática enfocada a las funcionalidades que un usuario en la silla de ruedas automatizada (liderando por el grupo de Investigación de Automática de la UAM) va a manejar. Para la realización de la aplicación se adaptó el modelo de lenguaje que para este idioma proporciona la SAPI (Interfaz de Programación de Aplicaciones de Voz) de Microsoft de manera que reconozca solo la gramática de interés. De esta manera se definen gramáticas que reconozcan comandos relacionados con el movimiento de la silla de ruedas, con órdenes de domótica y con la toma de signos vitales. La interfaz gráfica es diseñada para guiar al usuario en los comandos a pronunciar, el desplazamiento entre las ventanas de la aplicación también se puede controlar por voz así como el accionamiento de sus principales botones. Se proporciona así mismo la opción de desactivar o activar por voz el sistema de reconocimiento como medida de seguridad, importante cuando el usuario esta por ejemplo entablando una conversación con otra persona o lo que está pronunciando no va dirigido a la aplicación desarrollada. Otras funcionalidades como el envío por comandos de voz de correos electrónicos a destinatarios con su plantilla previamente almacenada por el usuario y la apertura por voz de programas instalados en el computador también son implementadas. Es de aclarar que este proyecto se realiza como trabajo de grado para optar al título de Magister en Mecatrónica y control de la Universidad Autónoma de Manizales y se encuentra inmerso dentro del macro proyecto de silla de ruedas automatizada que se trabaja al interior del grupo de Automática de la misma universidad.

El propósito de un sistema de reconocimiento del habla es tomar como entrada la forma de onda acústica de la voz humana y producir como salida una cadena de palabras equivalente [1]. 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; el decodificador utiliza tanto un modelo acústico como un modelo de lenguaje para generar finalmente la secuencia de palabras que tienen la máxima probabilidad de asemejarse a los vectores de características de entrada [2]. El modelo acústico es esencial para definir el comportamiento del sistema, este se obtiene con corpus de habla (ficheros de voz que contienen los datos de una amplia población de oradores con su correspondiente transcripción) de voces recogidas en el mismo idioma en el que se realizará el reconocimiento, mientras más robusto sea el corpus mejor será su desempeño. Si bien existen varias herramientas de software para realizar desarrollos con reconocimiento del habla, el hecho de que este proyecto es desarrollado para comandos en

español limita su escogencia y finalmente se opta por hacer el desarrollo con el SAPI de Microsoft que para este idioma ya tiene un desarrollo importante. Otras herramientas como “Julius” solo ponen a disposición modelos acústicos completos en japonés o en otros pocos idiomas principalmente el inglés.

Las pruebas para medir el desempeño del sistema de reconocimiento se realizan de manera discriminada por género y se desarrollan en tres ambientes con rangos de nivel de ruido que van desde los 35 dB(A) a 55 dB(A) para un ambiente tranquilo hasta un rango máximo de 73 dB(A) a 85 dB(A) para un ambiente ruidoso, rangos que parten de la actual legislación Colombiana sobre niveles máximos permisibles de ruido ambiental. Con la ayuda de matrices de confusión para interpretar los resultados, se obtienen los parámetros de eficiencia del reconocedor referentes a la exactitud, la sensibilidad, la especificidad, la precisión y la medida F1. Se resalta que el reconocimiento obtenido es independiente del hablante sin necesitar de los extensos entrenamientos previos que con otras herramientas se debe hacer.

1. REFERENTE CONTEXTUAL

1.1. ANTECEDENTES

La *Clasificación Internacional del Funcionamiento, de la Discapacidad y de la Salud* (CIF), define la discapacidad como un término genérico que engloba deficiencias, limitaciones de actividad y restricciones para la participación. La discapacidad denota los aspectos negativos de la interacción entre personas con un problema de salud (como parálisis cerebral, síndrome de Down o depresión) y factores personales y ambientales (como actitudes negativas, transporte y edificios públicos inaccesibles, y falta de apoyo social) que impiden su participación plena y efectiva en la sociedad en pie de igualdad con los demás [3].

Se estima que más de mil millones de personas viven en todo el mundo con alguna forma de discapacidad, o sea, alrededor del 15% de la población mundial (según las estimaciones de la población mundial en 2010); de ellas, casi 200 millones experimentan dificultades considerables en su funcionamiento. La *Encuesta Mundial de Salud* señala que, del total estimado de personas con discapacidad, 110 millones (2,2%) tienen dificultades muy significativas de funcionamiento, mientras que la *Carga Mundial de Morbilidad* cifra en 190 millones (3,8%) las personas con una “discapacidad grave” (el equivalente a la discapacidad asociada a afecciones tales como la tetraplejía, depresión grave o ceguera) [3].

En Colombia, según la Dirección de Censos y Demografía del DANE, a Marzo de 2010, en su registro continuo para la localización y caracterización de las personas con discapacidad, en la clasificación “según estructuras o funciones corporales que presentan alteraciones”, indica que existen 2.018.078 personas afectadas, y de estas 42.289 se registran en Caldas. De este registro nacional, 413.269 personas se encuentran clasificadas en alteración de “El movimiento del cuerpo, manos, brazos, piernas”; de las cuales 8.300 se encuentran en Caldas [4].

La discapacidad motora se asocia a daños en el sistema nervioso central o periférico. Entre sus causas, se mencionan accidentes de tránsito, laborales y caseros; lesiones personales con armas de fuego, armas blancas y minas antipersonas; así como enfermedades cerebrovasculares [5]. Las estadísticas norteamericanas reportan que cada año 300.000 a 400.000 personas sufren lesiones de la médula espinal presentándose con más frecuencia en los adolescentes y con una relación de cuatro hombres a una mujer [6]. En Colombia la situación es similar, tomando como ejemplo las estadísticas de la Unidad de Trauma del Hospital Universitario del Valle, se tiene que las secuelas y

las resultantes de traumas en la médula espinal, muestran que se están generando cada mes 5 a 6 pacientes parapléjicos y por lo menos un cuadripléjico, lo que implica que solo en éste hospital cada año se generan 60 personas parapléjicas y 12 cuadripléjicos con todas las implicaciones sociales y económicas que esto determina. Los accidentes de tránsito son la causa preponderante (49% del global) y por lesiones penetrantes las más frecuentes son causadas por arma de fuego [6].

Para aquellas personas con discapacidad motora se han venido investigando e implementando diferentes métodos que les permitan tener control sobre su silla de ruedas. Actualmente es común encontrar en el comercio sillas de ruedas eléctricas controladas por joystick [7] [8] [9]; y los estudios, especialmente, para pacientes que también carecen de la capacidad del habla, se han centrado en el uso de señales electromiográficas (EMG) tomadas en diversos músculos [10], el uso de señales electrooculográficas (EOG) con las que se puede detectar el movimiento de los ojos [11] y el uso de señales electroencefalográficas (EEG) que registran la actividad bioeléctrica cerebral [12]. De igual manera se ha trabajado el control de movimiento de una silla de ruedas por medio de la detección de la dirección del rostro [13] y del movimiento de la lengua [14]; y para pacientes que si se pueden comunicar oralmente, se trabaja también en el reconocimiento de comandos de voz [15] [16] [17] [18] [19] [20], tema de interés en el presente proyecto.

Particularmente, en el reconocimiento de comandos de voz, han sido varias las técnicas desarrolladas, como la DTW (Alineamiento temporal dinámico), cruce por cero, redes neuronales y HMM (Modelo Oculto de Markov), siendo esta última una de las más populares [2].

En la actualidad gracias a los avances en la tecnología y en la capacidad de procesamiento computacional, las implementaciones en el reconocimiento de comandos de voz, las cuales incluyen algunas de las técnicas mencionadas anteriormente, se están realizando por medio de Procesadores especializados de DSP (Procesamiento digital de señales) para tal fin, como lo es el "*DSK TMS320C6711"* de Texas Instruments, el cual trabaja calculando la energía, el cruce por cero y la desviación estándar de la palabra pronunciada [17]; el "*VR-Stamp"* basado en el procesador *RSC4128* de Sensory inc, el cual ofrece reconocimiento dependiente e independiente del hablante, Síntesis de Voz, verificación del hablante, escucha continua, grabación y reproducción, entre otras funciones [16] [21]; el kit de desarrollo para reconocimiento de voz "*Voice Direct 364* " también de Sensory inc, que emplea redes neuronales en el entrenamiento de las palabras o frases a reconocer [18] [22]; y el HM2007 voice recognition IC de la empresa Norteamericana HMC, que cuenta entre sus propiedades con análisis de voz, procesos de reconocimiento y funciones para sistemas de

control, éste puede reconocer hasta 40 palabras seleccionadas por el usuario cada una con longitud máxima de 0.96 segundos o 20 palabras cada una con longitud de hasta 1.92 segundos y puede entrenarse para trabajar en múltiples lenguajes [23] [24].

A nivel de Software, se encuentran también diversos toolkits de código abierto para reconocimiento de voz, como lo es “Julius” [25] [26] [27] [28] desarrollado por diferentes Instituciones de Japón, “CMU Sphinx” [29] [30] [31] desarrollado por la Universidad de Carnegie Mellon – EEUU, “ISIP” [32] [33] desarrollado por la Universidad del estado de Mississippi, y especializado en HMM está “HTK” [34] [35] [36] desarrollado por Cambridge University. De igual manera la compañía Microsoft distribuye una plataforma para el desarrollo del reconocimiento y síntesis de voz en su sistema operativo Windows, llamada SAPI (Interfaz de Programación de Aplicaciones de Voz) [37].

Es de anotar, que durante las últimas tres décadas, diversas comunidades interesadas en los estudios del habla, han contribuido a fomentar la mejora constante de las tecnologías de voz [2]. Entre sus aportes se encuentra el desarrollo compartido de corpus (ficheros de voz con su correspondiente transcripción) [38], herramientas de software libre [27] [31] [33] y fijación de directrices para llevar a cabo procesos de reconocimiento de voz. Lo cual ha logrado que el desarrollo de aplicaciones y el estudio e investigación de las tecnologías del habla sean accesibles aún para quienes no posean un conocimiento profundo en el tema.

Se destaca además los esfuerzos que las compañías Apple y Google han realizado en proveer sistemas de reconocimiento de voz cada vez más robustos para los dispositivos móviles.

Siri es el software de reconocimiento de voz para los productos de la empresa Apple, requiere acceso a Internet y dentro de los idiomas que actualmente entiende se encuentra el inglés, español, francés, alemán, japonés, mandarín, cantonés, italiano y coreano. Siri comienza el reconocimiento sin necesidad de realizar entrenamientos previos por parte del usuario, de manera que va mejorando la respuesta a medida que aprende las características del acento y la voz del usuario, empleando algoritmos de reconocimiento de voz para clasificar la voz como uno de los dialectos y acentos que comprende. A medida que cada vez más gente usa Siri y está expuesto a más variedades idiomáticas, su capacidad para reconocer dialectos y acentos mejora día a día, pretendiendo así funcionar cada vez mejor. Siri no trabaja con comandos concretos si no que entiende la forma natural de hablar y si necesita más información para completar una tarea, éste la pide. Siri permite usar la

voz para enviar y escribir mensajes, programar reuniones, realizar llamadas telefónicas, crear un contacto en la agenda**,** buscar respuesta a todo tipo de preguntas, dar información de calendario y en general para reemplazar la escritura por teclado y controlar las funciones del iPhone con la voz [39].

Google por su parte, incorpora Google Now como una interfaz de usuario de lenguaje natural con el que responde preguntas, hace recomendaciones y realiza acciones mediante la delegación de las solicitudes a un conjunto de servicios web, por lo cual al igual que Siri se hace necesaria una conexión a internet. Google Now se encuentra disponible en Android, iPhones y iPads; reconoce en más de 50 idiomas detectando automáticamente el idioma que se está hablando dependiendo de los lenguajes que el usuario haya habilitado y una vez reconoce lo que se quiso decir, el sistema contesta en el idioma que se habló. La aplicación ofrece sugerencias basándose en la ubicación del usuario y ofrece información en función de los hábitos de búsqueda y los contenidos de servicios Google. Con google now se pueden realizar acciones por voz como: fijar una alarma, establecer recordatorios y eventos, consultar la agenda, acceder a la cámara del dispositivo; llamar, enviar un correo o mensaje de texto a un contacto; escuchar emisoras, reproducir canciones y películas que se tengan almacenadas en Google play; realizar cualquier tipo de búsquedas a través de Google Chrome, consultar información general como la hora, el estado del tiempo, realizar conversión de unidades, traducir palabras entre idiomas, entre muchas otras funciones [40].

Un proyecto similar al propuesto en este documento, se desarrolló en la Universidad de Tottori en Japón en el 2007 “Voice controlled intelligent wheelchair” [41], quienes se apoyaron en la herramienta de software “Julian” (la cual es otra versión de Julius) [27] para controlar una silla de ruedas por comandos de voz en idioma Japonés. Ellos adaptaron la silla de ruedas con un computador portátil en donde el proceso de reconocimiento fue llevado a cabo, obteniendo una tasa de reconocimiento exitoso del 98.3% para los comandos de movimiento; Además de lo anterior, se encuentran diversos artículos de desarrollos en control de sillas de ruedas por comandos de voz utilizando procesadores especializados de DSP existentes en el mercado para tal fin [16] [17] [18] [23], con el que obtienen un reconocimiento de palabras aisladas limitado por la capacidad de memoria del DSP y se hace necesario un entrenamiento previo por cada usuario para efectuar el reconocimiento.

En Colombia, se encuentran al respecto trabajos como “Diseño y construcción de un módulo automático controlable por voz adaptable a una silla de ruedas convencional” de la Universidad Cooperativa de Colombia en el año 2009 [42], quienes implementaron el sistema de reconocimiento de voz sobre microcontroladores DSPIC’s y utilizaron una red neuronal artificial como técnica de

reconocimiento. El módulo reconoce 5 palabras del idioma español (adelante, atrás, izquierda, derecha y alto) y fue entrenado para identificar palabras pronunciadas por un único hablante, con el que obtuvieron un porcentaje de acierto en el reconocimiento del 90% para el promedio de las palabras pronunciadas; otro trabajo a mencionar es “Diseño e implementación de un prototipo de reconocimiento de voz basado en modelos ocultos de Markov para comandar el movimiento de una silla de ruedas en un ambiente controlado” de la Universidad Pedagógica y Tecnológica de Colombia en el año 2007 [43], cuyo algoritmo fue implementado en Matlab, detectando palabras aisladas en un vocabulario pequeño, dependiente del locutor y en un ambiente controlado, con el que obtuvieron una eficiencia en el algoritmo del 96.08%; similar al anterior se encuentra el proyecto “Implementación de una metodología para la detección de comandos de voz utilizando HMM” del Instituto Tecnológico Metropolitano, Medellín- Colombia, en el año 2012 [44], que también fue implementado en Matlab y permite la identificación de comandos de voz, con un diccionario reducido. Para éste construyeron una base de datos con los comandos: adelante, atrás, derecha, izquierda y pare. Los resultados obtenidos presentaron que la palabra con mayor cantidad de aciertos es adelante con un porcentaje de 98%, y una dispersión del 2,4%. Por otra parte la palabra que mayor dificultad presenta es pare con un porcentaje de acierto de 87% con una dispersión de 9,3%.

Con propósitos comerciales, en Colombia se ha creado la empresa “GAMMABIT Soluciones Tecnológicas”, conformada por un grupo de Ingenieros electrónicos de la Universidad de Los Llanos quienes lograron crear una empresa de desarrollo tecnológico con el apoyo de Parquesoft, Bavaria y la Gobernación del Meta, y quienes ganaron en el 2008 el programa ‘Destapa Futuro’ de Bavaria con el proyecto “sillas de ruedas que se impulsan con voz”, quienes pretenden sacar al mercado un dispositivo para el control de movimientos de sillas de ruedas eléctricas por comandos de voz [45].

1.2. FORMULACIÓN DEL PROBLEMA

¿Qué grado de desempeño se obtiene en el reconocimiento de comandos de voz en español, a través de la definición de un modelo de lenguaje para transmitir órdenes de control a una silla de ruedas e interactuar con algunas funciones del sistema operativo que administra la silla?

1.3. JUSTIFICACION

Como lo indica el informe mundial sobre la discapacidad, “En todo el mundo, las personas con discapacidad tienen peores resultados sanitarios y académicos, una menor participación económica y unas tasas de pobreza más altas que las personas sin discapacidad. En parte, ello es consecuencia de los obstáculos que entorpecen el acceso de las personas con discapacidad a servicios que muchos de nosotros consideramos obvios, en particular la salud, la educación, el empleo, el transporte, o la información. Esas dificultades se exacerban en las comunidades menos favorecidas” [3].

Conscientes de lo anterior y en la búsqueda por lograr disminuir en alguna medida las dificultades que en su desplazamiento deben afrontar las personas con discapacidad motora, el grupo de investigación en Automática de la Universidad Autónoma de Manizales UAM® con el apoyo del grupo de Diseño Mecánico y Desarrollo Industrial "Archytas" y del grupo de Diseño y Complejidad de la misma Universidad viene trabajando en el *proyecto integrador de “silla de ruedas automatizada”.* Dicho proyecto ha propuesto entre sus funciones el control del movimiento por comandos de voz, así como el accionamiento de dispositivos domóticos y de un módulo para transmitir automáticamente los signos vitales del paciente y enviar señales de pánico. Todo lo anterior con la posibilidad de ser controlado de igual manera por comandos de voz. Es entonces donde este proyecto de grado se encuentra dentro del marco del grupo de investigación de Automática de la Universidad Autónoma de Manizales, enfocándose al reconocimiento de comandos de voz en español con el que se pueda favorecer a pacientes que tienen la capacidad del habla pero que poseen dificultades graves o importantes para desarrollar actividades que requieren la utilización de movimientos finos y la destreza de los dedos de la mano, como lo es girar botones, perillas; la capacidad para alcanzar, tirar/halar o empujar objetos; y girar o torcer las manos o los brazos. Condiciones que no les permiten usar sus extremidades para hacer mover la silla de ruedas en la que se encuentran, ni accionar dispositivos de uso diario en su hogar o interactuar con un computador. Para esta clase de pacientes

el control de la silla de ruedas por comandos de voz es una opción cómoda, pues la voz sobresale como el medio de comunicación más natural y más usado para expresar lo que se desea.

Con el apoyo que ofrece la herramienta SAPI (Interfaz de Programación de Aplicaciones de Voz) de Microsoft donde tanto el modelo acústico como el modelo de lenguaje se encuentra funcionando para el idioma español, se pretende desarrollar una aplicación de software que presente un alto porcentaje de reconocimiento exitoso en este idioma, con una gramática definida según las necesidades de la aplicación, que sea independiente del hablante, sin necesidad de los extensos entrenamientos previos por cada usuario que se deben realizar bajo otras herramientas y con una interfaz amigable para el mismo, el cual aportará a mejorar la calidad de vida de personas con discapacidad motora y podrá ser reestructurado según necesidades de los grupos de investigación de la UAM®. Además por ser Windows 64 bits el sistema operativo en el que se va a trabajar, se asegura la compatibilidad de las herramientas ofrecidas por SAPI de Microsoft con este sistema operativo, y al desarrollar con .NET se hace posible interactuar con otros componentes realizados en distintos lenguajes de programación de manera sencilla [46].

1.4. OBJETIVOS

1.4.1 OBJETIVO GENERAL

Desarrollar una aplicación computacional para reconocimiento de comandos de voz en español e independiente del hablante que permita a un usuario comunicarse y transmitir órdenes de control a una silla de ruedas.

1.4.2 OBJETIVOS ESPECÍFICOS

- Desarrollar un modelo de lenguaje para el reconocimiento de comandos de voz en español orientado al control de una silla de ruedas eléctrica, aplicable en pacientes con limitaciones en la movilidad que puedan comunicarse oralmente.
- Implementar una interfaz iconográfica que permita a los usuarios de la silla de ruedas obtener una retroalimentación tanto escrita como auditiva del comando reconocido exitosamente, así como recordar los comandos a pronunciar.
- Determinar el nivel de confiabilidad o desempeño obtenido en la aplicación bajo distintos ambientes y según características del hablante.

2. REFERENTE TEÓRICO

2.1 Sistemas de reconocimiento de voz

Los sistemas de reconocimiento de voz, toman como entrada una forma de onda acústica y producen como salida una cadena de palabras [1]. Un sistema típico de reconocimiento de voz práctico consta de los componentes básicos que se muestran en el recuadro punteado de la Figura 1. Una interfaz de aplicaciones junto con el decodificador llega a resultados de reconocimiento que pueden utilizarse para adaptar otros componentes en el sistema [2].

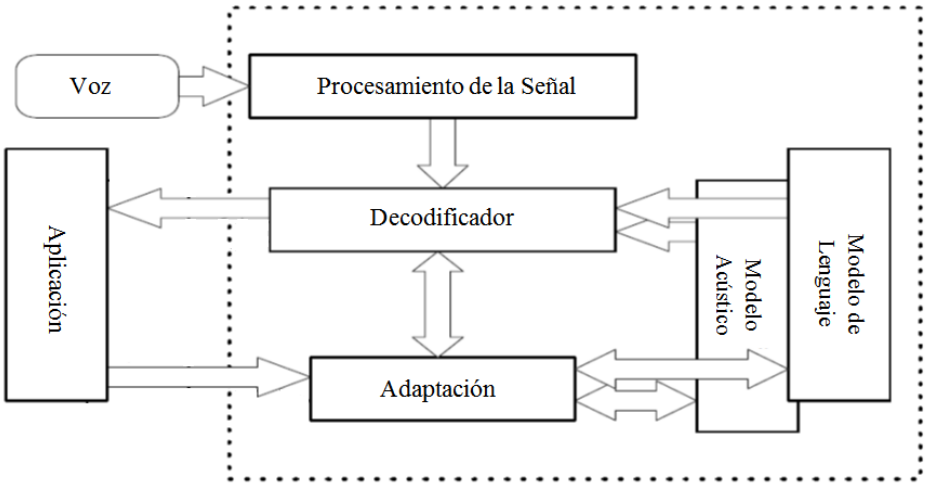

Figura 1: Arquitectura básica de un sistema de reconocimiento de voz. Adaptado de [2]

El modelo acústico se crea mediante la extracción de datos estadísticos de ficheros con voces, esta información estadística es una representación del sonido que forma cada palabra. Mientras más información de voces se tenga, el modelo acústico será más exacto [38]. Este modelo incluye también información acerca de la acústica, la fonética, el micrófono y la variabilidad del medio ambiente, género y diferencias dialectales entre los hablantes, etc [47].

Los modelos de lenguaje se refieren al conocimiento del sistema de lo que constituye una posible palabra, que palabras son probables de que co-ocurran, y en que secuencia. La semántica también puede ser necesaria para este modelo [47]. El modelo de lenguaje utilizado depende de la aplicación, puede ser un fichero de gramática que contiene conjuntos de combinaciones de palabras (en el caso de reconocimiento de comandos) que describe la posible sintaxis o patrones de las palabras en una tarea específica, siendo de esta manera un modelo basado en una gramática escrita; o un fichero que

contiene la probabilidad de que aparezcan ciertas palabras en un determinado orden (en el caso de aplicaciones de dictado) [38].

La señal de voz ingresa a un módulo de procesamiento de señales que extrae los vectores de características sobresalientes para ingresarlos al decodificador. El decodificador utiliza tanto el modelo acústico como el de lenguaje para generar la secuencia de palabras que tienen la máxima probabilidad a posteriori para el vector de entrada de características. Este puede también proporcionar información necesaria para que el componente de adaptación modifique ya sea el modelo acústico o el de lenguaje de manera que se puede obtener mejoras en el desempeño del sistema [47].

En las últimas décadas, los enfoques estadísticos en el manejo de los datos han logrado resultados alentadores [47], los cuales se basan por lo general en el modelado de la señal de voz usando algoritmos estadísticos bien definidos que pueden extraer automáticamente conocimiento a partir de los datos. El enfoque del manejo de los datos puede ser visto fundamentalmente como un problema de reconocimiento de patrones [47].

El decodificador de voz tiene entonces como objetivo, decodificar la señal acústica *X* en una secuencia de palabras $\hat{W}$, que en el caso ideal se aproxima a la secuencia de la palabra original *W*. Observar Figura 2.

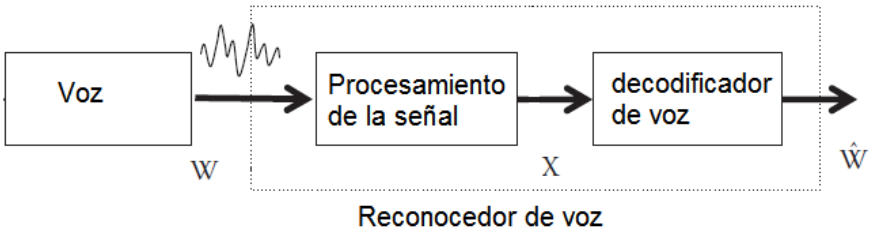

Figura 2: Modelo de un sistema típico de reconocimiento de voz. Adaptado de [2]

En la Figura 3 se observa un ejemplo de un diagrama de sistema genérico de reconocimiento de voz basado en modelos estadísticos, incluyendo el proceso de entrenamiento y decodificación.

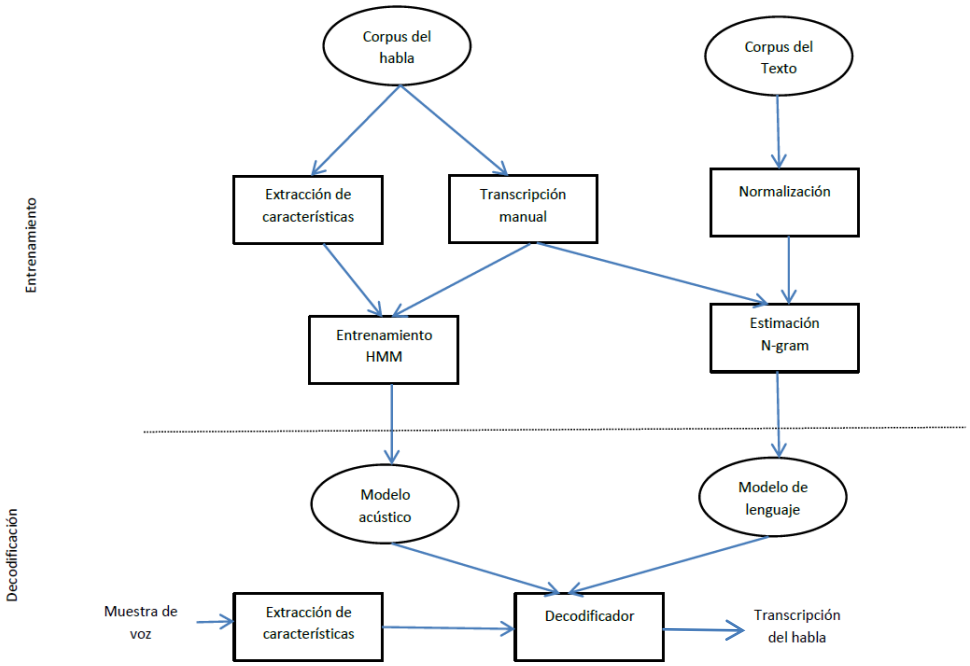

Figura 3: Ejemplo de un diagrama de sistema genérico de reconocimiento de voz basado en modelos estadísticos. Adaptado de [48]

2.2 Análisis de las características acústicas

2.2.1 Procesamiento digital de la señal - Codificación

El primer paso en el análisis de las características acústicas es la digitalización, donde la señal de voz continua es convertida en muestras discretas. En este proceso de convertir una señal análoga a una señal digital se pueden destacar tres pasos a: el muestreo, la retención y la cuantificación.

Una señal es muestreada al medir su amplitud en un momento particular; la tasa o frecuencia de muestreo se refiere al número de muestras tomadas por segundo. Para medir con precisión una onda, se necesitan al menos dos muestras por cada ciclo: una medida para la parte positiva de la onda y una medida para la parte negativa. Según el teorema de muestreo de Nyquist, para poder replicar con exactitud la forma de una onda es necesario que la frecuencia de muestreo sea superior al doble de la máxima frecuencia a muestrear. La mayor parte de la información en el habla humana se encuentra en frecuencias bastante inferiores a los 10000 Hz, por lo que una velocidad de muestreo de 20000 Hz sería necesaria para una completa exactitud. [1].

El nivel muestreado debe mantenerse constante hasta que se tome la siguiente muestra. Esto es necesario para que el ADC (Conversor análogo digital) disponga del suficiente tiempo como para procesar el valor muestreado. Esta operación de muestreo y retención genera una forma de onda "en escalera" que se aproxima a la forma de onda analógica de entrada [49], como se muestra en la Figura 4.

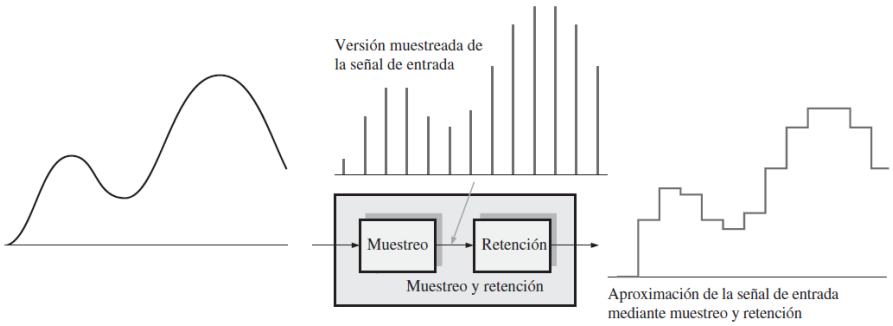

Figura 4: Operación de muestreo y retención. Tomado de [49]

Durante el proceso de cuantificación, el ADC convierte cada valor muestreado de la señal analógica en un código binario. Cuantos más bits se empleen para representar un valor muestreado, más precisa será la representación [49]. La Figura 5 ilustra la forma de onda de salida del bloque de muestreo y retención con dieciséis niveles de cuantificación. También se muestra la forma de onda analógica original como referencia. Los códigos binarios y los números de bits se han elegido arbitrariamente con propósitos de ilustración. La Figura 5 también muestra la forma de onda de salida del ADC, que representa los códigos binarios.

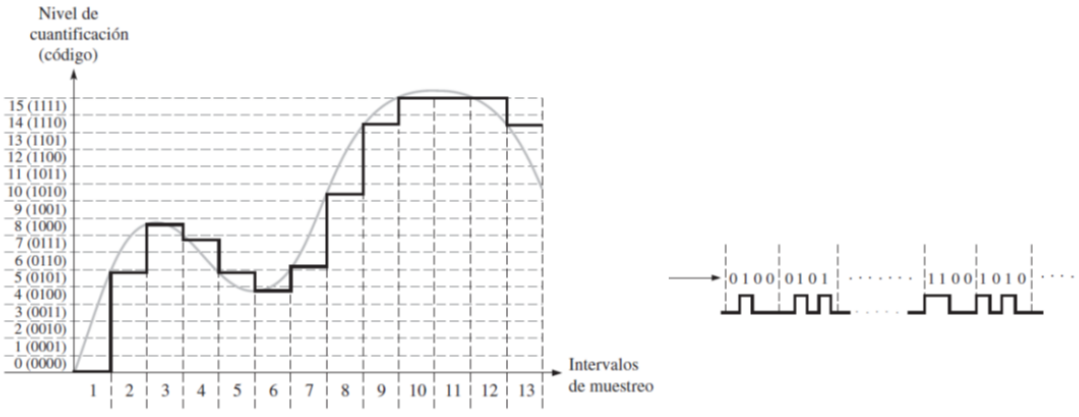

Figura 5: Forma de onda de salida del bloque de muestreo y retención con dieciséis niveles de cuantificación. La figura también muestra la forma de onda de salida del ADC, que representa los códigos binarios. Adaptado de [49]

Una vez que los datos se cuantifican, se almacenan en varios formatos. Un parámetro de estos formatos es la frecuencia de muestro y el tamaño de la muestra; la voz a través de teléfono es filtrada por la red de conmutación, y sólo las frecuencias de menos de 4000 Hz son transmitidos por los teléfonos, por lo tanto comúnmente se muestrea a 8 kHz y se almacena como muestras de 8 bits, mientras que los datos del micrófono a menudo se muestrea a 16 kHz y se almacena como muestras de 16 bits [1]. Otros parámetros son el número de canales (1 para mono, 2 para estéreo, 4 para el sonido cuadrafónico, etc.) y las técnicas de compresión que son la herramienta fundamental de la que se dispone para alcanzar el compromiso adecuado entre capacidad de almacenamiento y de procesamiento requeridas.

El Audio digital almacenado en los computadores (Windows WAV, Apple AIF, Sun AU, y SND, entre otros formatos) utiliza como su principal formato de codificación la PCM lineal (Modulación por codificación de pulsos) [47]. La PCM codifica cada muestra con un número fijo de bits que conformarán un tren de impulsos. Este tren de impulsos es una señal de alta frecuencia portadora de la señal analógica original.

2.2.2 Extracción de características

El siguiente paso es la extracción de características (también llamada parametrización), que tiene el objetivo de representar la señal de audio de una manera más compacta, tratando de eliminar la redundancia y reducir la variabilidad, mientras se mantiene la información lingüística importante [48]. En éste paso se transforma la forma de onda de entrada en una secuencia de vectores de características acústicas, en donde cada vector representa la información en una pequeña ventana de tiempo de la señal.

La mayoría de los sistemas de reconocimiento utilizan características cepstrales de corta duración basado ya sea en una transformada de Fourier o en un modelo de predicción lineal. Los parámetros Cepstrales son populares porque son una representación compacta, y están menos correlacionados que los componentes espectrales directos. Esto simplifica el proceso del modelo acústico mediante la reducción de la necesidad de modelar la dependencia característica. [48].

Si bien hay varios métodos para la extracción de características, uno de los más comunes en el reconocimiento de voz es el MFCC (coeficientes cepstrales en las frecuencias de Mel) [1]. En la Figura 6 se muestran los pasos para la extracción de los vectores de características por MFCC.

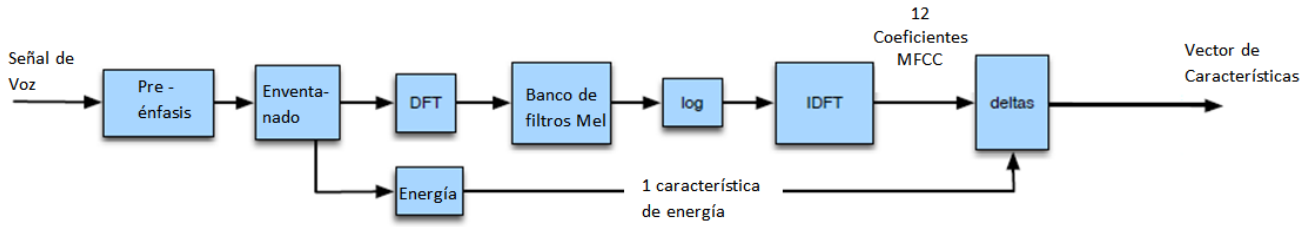

Figura 6: Pasos para la extracción de los vectores de características por MFCC de una forma de onda digitalizada cuantificada. Adaptada de [1]

El proceso para la extracción del vector de características MFCC mostrado en la Figura 6 se describe en los apartados a continuación.

2.2.2.1 Preénfasis

La primera etapa en la extracción de características MFCC es aumentar la cantidad de energía en las frecuencias altas. Resulta que en el espectro de segmentos de voz como las vocales, hay más energía en las frecuencias bajas que en las frecuencias más altas. Esta caída de la energía a través de las frecuencias (que es llamado inclinación espectral) es causada por la naturaleza del impulso glotal [1]. En el proceso de generación de la voz, el sonido inicial proviene de la vibración de las cuerdas vocales conocida como vibración glotal, en donde, el efecto sonoro se genera por la rápida apertura y cierre de las cuerdas vocales conjuntamente con el flujo de aire emitido desde los pulmones. La abertura entre las dos membranas que conforman las cuerdas vocales se denomina glotis.

Al Aumentar la energía de alta frecuencia se hace que la información de estos formantes superiores sea más asequible a los modelos acústicos y mejora la exactitud en la detección de los fonemas. Éste paso se realiza con un filtro pasa altas de primer orden. Un Fonema es un sonido del habla; los fonemas están representados con símbolos fonéticos que guardan alguna semejanza de una letra en un lenguaje alfabético como el español [1].

2.2.2.2 Ventaneo

La señal de voz es un proceso aleatorio y no estacionario, lo que significa que sus propiedades estadísticas no son constantes a través del tiempo debido a las limitaciones físicas de la velocidad a la que los articuladores pueden moverse. Esto supone un inconveniente a la hora de analizarla. No obstante, es posible salvar este problema si se tiene en cuenta que a corto plazo de tiempo (del orden de 10 ms a 20 ms) la señal es casi-estacionaria [48]. Se extrae esta porción del habla aproximadamente estacionaria mediante el uso de una ventana que no es cero dentro de una región y es cero en la otra parte, se ejecuta la ventana a través de la señal de voz y se extrae la forma de onda dentro de esta [1].

Para extraer la señal se multiplica el valor de la señal en el momento *n*, *s[n],* con el valor de la ventana en el tiempo *n*, *w[n]*:

$$y[n] = w[n]s[n] \quad (1)$$

Con el fin de evitar componentes falsos de alta frecuencia en el espectro debido a discontinuidades causadas por la división en ventanas de la señal y que pueden crear problemas cuando se aplique el análisis de Fourier, es común utilizar una ventana cónica como la ventana Hamming [48] [1]. La cual reduce los valores de la señal hacia cero en los límites de la ventana, evitando discontinuidades.

Así mismo, para mantener la continuidad de la información de la señal, se suele realizar el ventaneo con bloques de muestras solapados entre sí, de tal manera que no se pierde información en la transición entre ventanas. Generalmente el solapamiento se lleva a cabo con un desplazamiento entre ventanas de 10 ms, obteniéndose coeficientes MFCC cada 10 ms [48].

La Figura 7 muestra el ventaneo de una señal senoidal pura con una ventana rectangular y una ventana Hamming.

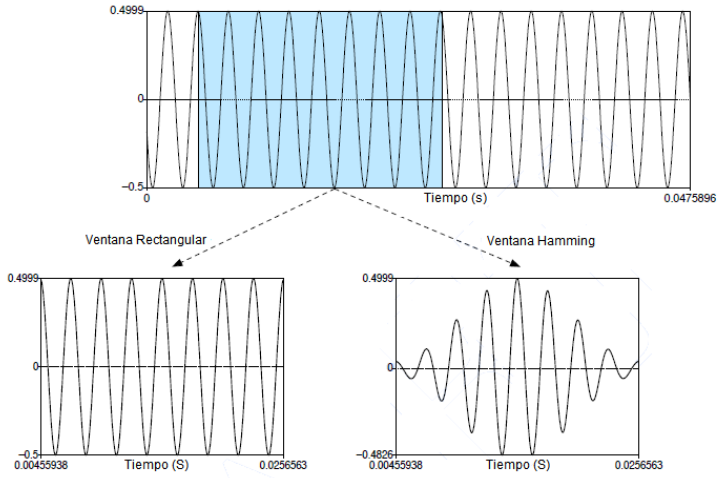

Figura 7: Ventaneo de una porción de una onda sinusoidal pura con la ventana rectangular y la ventana de Hamming. Adaptado de [1]

2.2.2.3 Transformada Discreta de Fourier (DFT)

El siguiente paso es extraer la información espectral para la señal de ventaneo; se necesita saber entonces cuánta energía contiene la señal en diferentes bandas de frecuencia. La herramienta para extraer información espectral de las bandas de frecuencia discretas para una señal de tiempo discreto (muestreada) es la Transformada Discreta de Fourier (DFT) [1].

La entrada para la DFT es la señal ventaneada x[n]...x[m], y la salida para cada una de las N bandas de frecuencia discreta es un número complejo X[k] que representa la magnitud y fase de dicho componente de frecuencia en la señal original. Al graficar la magnitud contra la frecuencia, se puede visualizar el espectro de la señal. Éste espectro es útil ya que los picos espectrales que son fácilmente visibles en un espectro son característicos de los diferentes fonos; los fonos tienen una "firma" espectral característica. Se puede detectar la firma característica de los diferentes fonos mirando el espectro de una forma de onda. Un algoritmo común usado para calcular la DFT es la Transformada Rápida de Fourier (FFT) [1].

2.2.2.4 Banco de filtros Mel y Log

Los resultados de la FFT dan información acerca de la cantidad de energía en cada banda de frecuencia. Sin embargo, la audición humana no es igualmente sensible a todas las bandas de frecuencia. Ésta es menos sensible a frecuencias altas, aproximadamente por encima de 1000 Hz. Al modelar esta propiedad de la audición humana durante la extracción de características se mejora el rendimiento de reconocimiento de voz [1].

El tono de un sonido es la sensación mental o correlación perceptual de la frecuencia fundamental; en general, si un sonido tiene una alta frecuencia fundamental, el humano lo percibe como que tiene un tono alto. Como la relación no es lineal, el oído humano tiene diferentes agudezas para diferentes frecuencias. El modelo psicoacústico de las escalas de la percepción del tono que utiliza el MFCC (coeficientes cepstrales en las frecuencias de Mel) es la escala Mel [1]. La escala Mel se aproxima a la resolución de frecuencia del sistema auditivo humano, siendo lineal en el rango de baja frecuencia (por debajo de 1000 Hz) y logarítmica por encima de 1000 Hz [48]. Un mel es entonces una unidad del tono. Por definición, pares de sonidos que están perceptualmente equidistante en tono son separados por un igual número de mels [1]. Para convertir *f* Hz en *m* Mels se emplea:

$$Mel = 1127 \ln\left(1 + \frac{f}{700}\right) \quad (2)$$

En los cálculos de MFCC se implementa un banco de filtros triangulares (de área unidad) que colectan la energía de cada banda de frecuencia. Estos triángulos están espaciados de acuerdo con la escala de frecuencias Mel, en donde se tienen 10 filtros espaciados linealmente por debajo de 1000 Hz y los filtros restantes se extienden de forma logarítmica por encima de los 1000 Hz [1].

Posteriormente se saca el logaritmo de los parámetros de salida del banco de filtros triangulares (valores del espectro Mel). En general, la respuesta humana al nivel de la señal sonora es logarítmica; además utilizando un logaritmo se logra que la estimación de características sea menos sensible a las variaciones en la entrada como lo es las variaciones de potencia debido a la boca del locutor moviéndose más cerca o más lejos del micrófono) [1].

2.2.2.5 Parámetros cepstrales - Transformada Inversa de Fourier Discreta

El inconveniente de trabajar en el dominio espectral, es que los espectros de los filtros en las bandas adyacentes presentan un alto grado de correlación, originando coeficientes espectrales estadísticamente muy dependientes entre ellos. Por esta razón, el siguiente paso en la extracción de características MFCC es el cálculo del cepstro. Los parámetros cepstrales son obtenidos tomando la transformada inversa del logaritmo de los parámetros del banco de filtros triangulares. La unidad correcta de un cepstro es la muestra [1].

El cepstro tiene ventajas de procesamiento útiles y mejora significativamente el desempeño en el reconocimiento del fono. La señal del habla es creada cuando una forma de onda de la fuente glotal de una frecuencia fundamental en particular pasa a través del tracto vocal, que debido a su forma tiene una característica de filtrado particular. Pero muchas características de la fuente glotal (su frecuencia fundamental, los detalles del pulso glotal, etc) no son importantes para distinguir los diferentes fonos. En lugar de ello, la información más útil para la detección de fono es el filtro, que se refiere a la posición exacta del tracto vocal. Si se conoce la forma del tracto vocal, se sabe que fono se está produciendo. El cepstro es un medio para separar la fuente y el filtro y mostrar solamente el filtro del tracto vocal, lo cual resulta importante en la detección de fonos [1].

Si se está interesado en detectar fonos, se puede hacer uso de los valores cepstrales más bajos. Si se está interesado en detectar tonos, se puede hacer uso de los valores cepstrales más altos. Para la extracción MFCC, generalmente se toman los 12 primeros valores cepstrales. Estos 12 coeficientes representarán únicamente la información sobre el filtro de tracto vocal, limpiamente separada de la información sobre la fuente glotal [1].

2.2.2.6 Deltas y Energía

La energía se correlaciona con la identidad del fonema siendo así una señal útil para la detección de este último, por ejemplo las vocales tienen más energía que las pausas. La energía en una trama es la suma en el tiempo de la potencia de las muestras en la trama; Así, para una señal *x* en una ventana de muestra de tiempo *t1* a una muestra de tiempo *t2*, la energía es [1]:

$$Energía = \sum_{t=t1}^{t2} x^2\ [t] \qquad (3)$$

De la extracción del Cepstro con la transformada inversa de Fourier discreta mencionada en el paso anterior resultaron 12 coeficientes cepstrales por cada trama. La adición de la característica 13 representa la energía de la trama [1].

Con el fin de captar la naturaleza dinámica de la señal de voz, es común aumentar el vector de características con parámetros delta [48]. De ésta manera, además de los MFCC, si se desea obtener otra información, como por ejemplo, la coarticulación de los fonemas, se deben incluir los coeficientes MFCC-Delta y los MFCC-Doble Delta que representan la evolución temporal de los fonemas en su transición a otros fonemas, y en general permiten tener en cuenta la variabilidad de un interlocutor a la hora de hablar. Los MFCC-Delta se le conocen como coeficientes de velocidad, ya que miden la variación de los coeficientes MFCC sobre un instante de tiempo. De la misma forma, a los MFCC-Doble Delta se les denomina coeficientes de aceleración, pues representan la variación de los MFCC-Delta sobre un instante de tiempo.

Por cada una de las 13 características (12 características cepstrales mas la energía) se adhiere un coeficiente de velocidad y otro de aceleración [1]. Sumando todas las características anteriores, finalmente se obtiene un vector con 39 características MFCC por cada trama, como se referencia en la Tabla 1.

12	Coeficientes cepstrales
12	Coeficientes cepstrales delta
12	Coeficientes cepstrales doble delta
1	Coeficiente de energía
1	Coeficiente de energía delta
1	Coeficiente de energía doble delta
39	Características MFCC

Tabla 1: Características MFCC. Tomado de [1]

2.3 Modelo Acústico y Modelo de lenguaje

Si consideramos una señal acústica *A*, el proceso de reconocimiento en una aproximación estocástica consiste en calcular la probabilidad *P(WIA)* de que la secuencia de palabras o frase *W* corresponda a la señal acústica *A*, y encontrar la secuencia de palabras con mayor probabilidad [50]. Usando la regla de Bayes, esta probabilidad puede escribirse como

$$P(W|A) = \frac{P(W)P(A|W)}{P(A)} \quad (4)$$

donde *P(W)* es la probabilidad de la secuencia de palabras *W*, *P(AIW)* es la probabilidad de la señal acústica *A* dada una secuencia de palabras *W*, y *P(A)* es la probabilidad de la señal acústica. Por tanto, es necesario tener en cuenta *P(AIW),* que es el modelo acústico, y *P(W),* que es el modelo de lenguaje. Ambos modelos pueden representarse mediante modelos de Markov [50].

La división del modelo acústico y del modelo de lenguaje se puede describir de manera concisa por la ecuación fundamental del reconocimiento de voz estadístico:

$$\widehat{W} = arg_w \, max \, P(W|A) = arg_w \, max \frac{P(W)P(A|W)}{P(A)} \quad (5)$$

Para la observación acústica dada, el objetivo del reconocimiento de voz es encontrar la correspondiente secuencia de palabras $\hat{W} = w_1 w_2 \ldots w_m$ que tenga la máxima probabilidad posterior *P(W|A)* tal como se expresa con la ecuación (**5**) Dado que la maximización de ésta ecuación se lleva a cabo con la observación de *A* fijo, la maximización es equivalente a la maximización del numerador:

$$\widehat{W} = arg_w \max P(W)P(A|W) \quad (6)$$

Donde *P(W) y P(A|W)* constituyen las cantidades probabilísticas calculadas por los componentes del modelo del lenguaje y del modelo acústico, respectivamente, de los sistemas de reconocimiento de voz. El desafío práctico es cómo construir modelos acústicos precisos, *P(A|W),* y modelos de lenguaje, *P(W)*, que puedan realmente reflejar el lenguaje hablado para ser reconocido [2].

2.3.1 Modelo Acústico

El modelo acústico juega un rol fundamental para mejorar la precisión en el reconocimiento automático del habla. Hay un número de factores conocidos que determinan la precisión de los sistemas de reconocimiento, dos de los más notables son las variaciones del contexto y variaciones de los hablantes. No es descabellado exponer que el modelo acústico es la parte central de cualquier sistema de reconocimiento del habla [2], éste incluye la representación del conocimiento acerca de la acústica, la fonética, el micrófono y la variabilidad del medio ambiente, género y diferencias dialectales entre los hablantes, etc [47].

El modelo acústico típicamente se refiere al proceso de establecer representaciones estadísticas para las secuencias del vector de características calculado de la forma de onda del habla de ficheros con voces [2] [48]. Mientras más información de voces se tenga, el modelo acústico será más exacto [38]. Un modelo acústico independiente del hablante es posible tomando varios datos de audio para el entrenamiento. Según foro oficial de la página “Julius” [27], con un mínimo de 20 – 30 grabaciones de diferentes personas en las que queden registrados todos los comandos deseados y más de una vez en lo posible, se consiguen resultados tangibles. Se debe tener en cuenta que la calidad de las grabaciones afecta grandemente el desempeño en el reconocimiento, por lo que estas deben tener la menor distorsión y menor ruido posible, así como el volumen debe ser el apropiado para evitar el recorte en voz alta [28].

El Modelo acústico también abarca “el modelo de pronunciación”, que describe cómo una secuencia o multi-secuencia de unidades fundamentales del habla (tales como fonos o rasgos fonéticos) son usados para representar unidades de habla más grandes como palabras o frases que son el objeto del reconocimiento de voz. El Modelo acústico también puede incluir el uso de información de realimentación desde el reconocedor para formar de nuevo los vectores de características del habla hacia el logro de la robustez frente al ruido en el reconocimiento [2].

2.3.1.1 Modelos ocultos de Markov

Los modelos ocultos de Markov (HMM) son uno de los tipos más usados de modelos acústicos [48] [2] [1], con el que no sólo se puede proporcionar una forma eficiente de construir modelos paramétricos, sino que también puede incorporar el principio de programación dinámica en su núcleo, útil para un patrón de segmentación unificada y clasificación de patrones de secuencias de datos variables en el tiempo [47]. Además el HMM ha podido incorporar en un modelo común tanto el modelo acústico de bajo nivel (unidades lingüísticas y silencios) como el del lenguaje de alto nivel (gramática), lo que ha permitido que se pueda realizar al mismo tiempo la segmentación y el reconocimiento de las unidades lingüísticas sin necesidad de emplear un detector de silencios como sí lo necesitan otros métodos como los DTW (alineamiento temporal dinámico) y las redes neuronales. A su vez, esto ha dado origen al éxito de los HMM en el reconocimiento de habla continua con grandes vocabularios empleando como unidades lingüísticas fonemas o trifonemas [51]. El HMM es por lo tanto un poderoso método estadístico de caracterización de las muestras de datos observados de una serie de tiempo discreto [47].

El HMM se utiliza entonces para modelar la producción de vectores de características del habla en dos pasos. En primer lugar una cadena de Markov se utiliza para generar una secuencia de estados, y como segundo, vectores de voz se dibujan usando una función de densidad de probabilidad (PDF) asociada a cada estado. La cadena de Markov se describe por el número de estados y las probabilidades de transición entre estados [48]. Actualmente, las unidades acústicas elementales más utilizadas en sistemas para el reconocimiento continúo del habla para vocabulario amplio se basan en el fono, donde cada fono está representado por una cadena de Markov con un pequeño número de estados [48].

La Figura 8 (a) muestra una cadena de Markov para asignar una probabilidad a una secuencia de fenómenos meteorológicos, donde el vocabulario consta de Hot (caliente), Cold (frío), y Warm (cálido), aquí las unidades lingüísticas consideradas no son fonemas o trifonemas sino palabras. La Figura 8 (b) muestra otro ejemplo sencillo de una cadena de Markov para asignar una probabilidad a una secuencia de palabras $w_1...w_n$. Dados los modelos de la Figura 8, podemos asignar una probabilidad a cualquier secuencia del vocabulario [1].

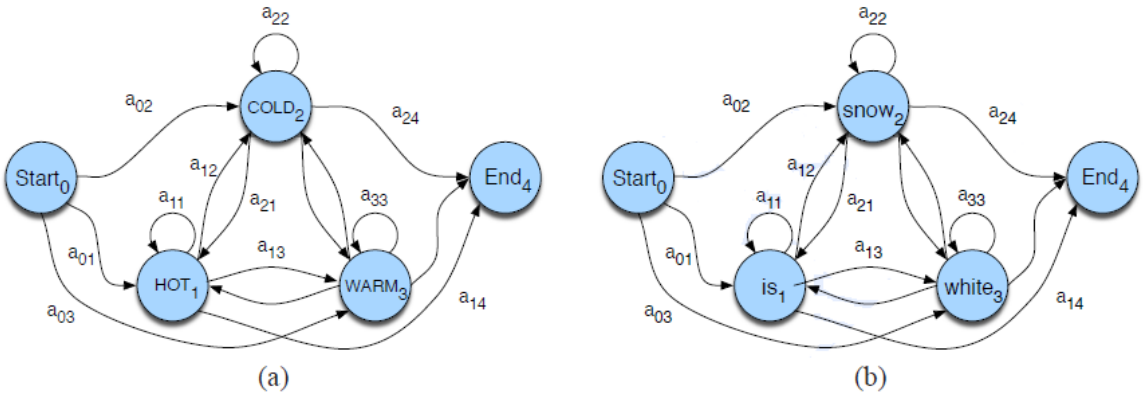

Figura 8: Cadena de Markov para para un vocabulario relacionado con el clima (a) y para una secuencia de palabras (b). Tomado de [1]

La técnica que emplea los HMM, está compuesta por un conjunto de estados conectados por transiciones, el cual comienza en un estado inicial designado. En cada paso de tiempo discreto, una transición se toma dentro de un nuevo estado y luego un símbolo de salida se genera en él. La elección de la transición y el símbolo de salida son ambos aleatorios, gobernados por distribuciones de probabilidad [52]. Las unidades lingüísticas (y los silencios) son de esta manera HMM definidos por sus estados *q*, sus probabilidades de transición entre estados $a_{i,j}$ y sus probabilidades de emisión

p(x|q) de la observación *x* dado el estado *q* que se refieren como probabilidades de observación. Las probabilidades de transición entre las palabras vienen dadas por el modelo o gramática del lenguaje. Cada estado suele representar un segmento de señal cuasiestacionario (casi un fonema). La topología de los HMMs de cada palabra es normalmente hacia delante [51]. Los HMM pueden ser considerados como una caja negra, donde la secuencia de símbolos de salida generados a través del tiempo es observable, pero la secuencia de estados visitados a través del tiempo se oculta a la vista. Cuando un HMM se aplica al reconocimiento del habla, los estados se interpretan como modelos acústicos, indicando que sonidos son más probables a ser escuchados durante sus correspondientes segmentos de voz; mientras que las transiciones proporcionan restricciones temporales, indicando como los estados pueden seguir uno al otro en secuencia. Como el habla siempre va hacia adelante en el tiempo, las transiciones en una aplicación de voz siempre irán hacia adelante (o haciendo un auto-lazo que permite a un estado tener una duración arbitraria). Los estados y transiciones en un HMM pueden ser estructurados jerárquicamente, con el fin de representar fonemas, palabras y oraciones [52].

Un Modelo Oculto de Markov de primer orden asume dos supuestos simplificadores. Primero, la probabilidad de estar en un estado en particular depende únicamente del estado anterior:

$$\textit{Suposición de Markov}: \quad P(qi|q1 \ldots qi-1) = P(qi|qi-1) \tag{7}$$

Segundo, la probabilidad de una observación de salida *xi* depende sólo del estado actual que produjo la observación *qi* y no de cualquier otro estado o cualquier otra observación, es decir, las observaciones son independientes entre sí conocido el estado:

$$\textit{Independencia de salida}: \quad P(xi|q1 \ldots qi, \ldots, qT, x1, \ldots, xi, \ldots, xT) = P(xi|qi) \tag{8}$$

Propiamente hablando, dado un n-estado de HMM con vector de parámetros *λ*, el proceso estocástico HMM es descrito por la siguiente función de densidad de probabilidad conjunta $f(x, s|\lambda)$ de la señal observada $x = (x_1, \ldots, X_T)$ y la secuencia de estado no observada $s = (s_0, \ldots, s_T)$,

$$f(x, s|\lambda) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} f(x_t|s_t) \tag{9}$$

Donde π_i es la probabilidad inicial del estado *i*, a_{ij} es la probabilidad de transición del estado *i* al estado *j*, y $f(\bullet|s)$ es la PDF emitida asociada a cada estado *s* [48].

2.3.1.2 Otros métodos: Alineamiento Temporal Dinámico y Redes Neuronales

El alineamiento temporal dinámico (DTW), el cual hace parte de los algoritmos conocidos como de programación dinámica, ha sido ampliamente utilizado para obtener la distorsión total entre dos muestras de locución [47]. Esta técnica alinea los sonidos registrados en dos patrones distintos de una misma palabra, para garantizar así una comparación razonable de los datos. DTW tiene su punto de partida en el ajuste de plantillas (template matching), para llevar a cabo la prueba de ajuste de estas es indispensable que cada palabra se encuentre alineada en el tiempo con la plantilla que se encuentra en observación. Las cadenas de vectores extraídos de la frase de prueba y de la frase de referencia (template) se colocan sobre los ejes *x*, *y*, respectivamente [53]; Se calcula luego una ruta de alineación óptima para cada plantilla de la palabra de referencia, y la de menor puntuación acumulada es considerada como la mejor opción para la muestra de voz desconocida [52]. La ventaja de la programación dinámica utilizada por DTW radica en el hecho de que una vez un sub-problema se resuelve, el resultado parcial puede ser almacenado y no necesita ser recalculado, este principio es muy importante en la construcción de sistemas prácticos de lenguaje hablado. El reconocimiento de voz basado en DTW es fácil de implementar y resulta efectivo al trabajar con vocabularios pequeños. La programación dinámica puede temporalmente alinear patrones para tener en cuenta las diferencias en la tasa del habla al tomar varias muestras de la misma palabra pronunciadas por el mismo orador. Sin embargo, no puede obtener una plantilla promedio por cada patrón a partir de una gran cantidad de muestras de entrenamiento, lo que le da una desventaja al momento de caracterizar variaciones entre las diferentes expresiones, ya que para ello típicamente se requiere múltiples referencias de locutores [47].

Por su parte, las redes neuronales artificiales poseen una estructura que se asemeja a la transmisión y procesamiento de señales en las neuronas biológicas. La red neuronal artificial se compone de un número potencialmente elevado de elementos de procesamiento simple (llamadas unidades, nodos, o neuronas) quienes se influyen unos a otros en la conducta a través de una red de pesos de conexión excitadores o inhibidores. En ella, el *set* de entrenamiento consiste en patrones de valores que son asignados a unidades específicas de entrada y/o salida, entonces, una regla de aprendizaje modifica el poder de los pesos para que la red se aprenda gradualmente el set de entrenamiento. Estas redes son particularmente útiles en la extracción de características y el reconocimiento de patrones, así como en los sistemas de toma de decisiones. Por ser el reconocimiento del habla básicamente un

problema de reconocimiento de patrones, muchas investigaciones al respecto han aplicado redes neuronales, alcanzando un cierto éxito en el nivel de reconocimiento de palabras [52].

2.3.1.3 Entrenamiento

El entrenamiento estima los parámetros que caracterizan a los HMMs, previamente fijada la topología (número de estados, enlaces, etc). Consiste en disponer de múltiples representaciones acústicas (conjunto de entrenamiento) del sistema a modelar y a partir de ellas estimar los valores $a_{i,j}$ y $p(x|q)$ que mejor representen al conjunto de entrenamiento y por lo tanto del sistema a modelar [51]. El algoritmo más comúnmente empleado para estimar estos valores es el de Baum-Welch [48] [51] [7], el cual es un algoritmo iterativo tipo EM (maximización de la esperanza). Este algoritmo garantiza que la probabilidad de los datos de entrenamiento dados los modelos incremente en cada iteración. Algunos detalles de implementación, tales como un procedimiento de inicialización adecuada y el uso de restricciones en los valores de los parámetros pueden ser bastante importantes [48].

Dado que el objetivo del entrenamiento es encontrar el mejor modelo a considerar para los datos observados, el desempeño del reconocedor es críticamente dependiente sobre la representatividad de los datos de entrenamiento. La Independencia del hablante se obtiene mediante la estimación de los parámetros de los modelos acústicos en grandes corpus de habla que contienen los datos de una amplia población de oradores. Existen diferencias sustanciales en el habla de interlocutores masculinos y femeninos que surge de las diferencias anatómicas (en promedio las mujeres tienen una longitud del tracto vocal más corto que resulta en frecuencias de formantes superiores, así como en una mayor frecuencia fundamental). Por tanto, es práctica común utilizar modelos separados para hablantes femeninos y masculinos a fin de mejorar el rendimiento del reconocimiento, que a su vez requiere la identificación automática del género [48].

2.3.2 Modelo de Lenguaje

El rol del modelo de lenguaje en el reconocimiento de voz es proporcionar el valor $P(W)$ en la ecuación fundamental de reconocimiento de voz de la Ecuación (**6**) [2]. A través de dicho modelo se obtienen las probabilidades a priori de las secuencias de palabras a reconocer. Para estimar este valor para secuencias de cualquier longitud se necesitaría una gran cantidad de datos por lo que se debe acudir a aproximaciones. Las aproximaciones que están más extendidas son las basadas en N-

Gramas. En estos tipos de modelos de lenguaje la probabilidad de aparición de una palabra únicamente depende de un número reducido de palabras que la preceden [54]. En un modelo 2-grama (comúnmente llamado bigrama), la probabilidad de una palabra, dada la palabra anterior, se calcula como la frecuencia de secuencias de dos palabras, como por ejemplo “mover adelante” o “tomar presión”; En un modelo 3-grama (comúnmente llamado trigrama), se calcula la probabilidad de una palabra, dadas las dos palabras anteriores como lo son la secuencia de palabras “prender luz cocina” o “enviar correo electrónico”. Tales modelos estadísticos de secuencias de palabras también se llaman modelos de lenguaje. Por lo anterior, estimadores como los N-gramas que asignan una probabilidad condicional a posibles próximas palabras se pueden utilizar para asignar una probabilidad conjunta para una frase entera [54], [1]. La terminología de asociar el valor *N* de un modelo N-grama con su orden, proviene de los modelos de Markov, en donde un modelo N-grama puede ser interpretado como un modelo de Markov de orden *N-1* [2].

La distribución de probabilidad *P (W)* sobre la cadena de palabras *W*, puede descomponerse como:

$$P(W) = P(w_1, w_2, \dots, w_n)$$

$$= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_n|w_1, w_2, \dots, w_{N-1})$$

$$= \prod_{i=1}^{n} P(w_i|w_1, w_2, \dots, w_{i-1}) \qquad (10)$$

Donde $P(w_i|w_1, w_2, \dots, w_{i-1})$ es la probabilidad de que w_i le siga a la secuencia de palabras $w_1, w_2, \dots, w_{i-1}$ que le fueron presentadas previamente [2]. Los valores que intervienen en ésta ecuación se estiman usando ejemplos obtenidos a partir de bases de datos de textos.

Así, en un modelo bigrama para estimar $P(w_i|w_{i-1})$, el cual se refiere a la frecuencia con la que la palabra w_i ocurre dada la última palabra w_{i-1}, se realiza mediante un conteo de con qué frecuencia la secuencia $P(w_i|w_{i-1})$ ocurre en algún texto y se normaliza el conteo por el número de veces que w_{i-1} ocurre [2].

El trigrama se puede estimar mediante la observación de las frecuencias o recuentos del par de palabras $C(w_{i-2}, w_{i-1})$ y la tripleta $C(w_{i-2}, w_{i-1}, w_i)$ como sigue [2]:

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \quad (11)$$

Los N-gramas deben entrenarse con grandes bases de datos (corpus), de manera que entre mayor es el valor de N, más grande debe ser el corpus. Si la base de datos no es lo suficientemente grande y el número de palabras del vocabulario es alto, muchas sucesiones de palabras existentes de hecho no aparecerán y el modelo, especialmente en el caso de los trigramas, tendrá muchas probabilidades nulas. La ventaja de los modelos basados en palabras es que pierden menos información sintáctica y semántica. Además se entrenan de una manera muy simple, ya que el texto no necesita ningún etiquetado gramatical inicial. Sin embargo, la cantidad de datos necesaria para entrenar el modelo es muy grande, especialmente en el caso del trigrama. Cuando se usan categorías gramaticales, el texto deber ser etiquetado pero puede ser más corto. Además, si una nueva palabra es introducida en el diccionario puede heredar las probabilidades calculadas anteriormente para las palabras de la misma categoría gramatical [50].

La mejor métrica para evaluar un modelo de lenguaje es la tasa de error de reconocimiento de palabras, lo que requiere la participación de un sistema de reconocimiento de voz. Ésta es llamada evaluación extrínseca o evaluación en vivo. Desafortunadamente dicha evaluación es a menudo muy costosa; la evaluación de un conjunto de pruebas de un amplio vocabulario a reconocer, por ejemplo, toma horas o incluso días. Por otro lado, una métrica de evaluación intrínseca es una que mide la calidad de un modelo independiente de cualquier aplicación. La perplejidad es la métrica de evaluación intrínseca más común para modelos de lenguaje N-grama [1]. Ésta es una medida derivada de la entropía cruzada. Los métodos basados en entropía tratan de encontrar la red cuya entropía cruzada con los datos sea mínima. La entropía se puede considerar como una forma de medir el grado de dependencia entre variables, y en ese sentido estos métodos lo que hacen es buscar configuraciones que favorezcan la presencia de conexiones entre variables que manifiesten un alto grado de dependencia [55].

Dado un modelo de lenguaje que asigna la probabilidad *P(W)* a una secuencia de palabras *W*, se puede derivar un algoritmo de compresión que codifica el texto *W* usando - $log_2 P(T)$ bits. La entropía cruzada (la cual mide la media de bits necesarios para identificar un evento de un conjunto de posibilidades) *H(W)* de un modelo $P(w_i|w_{i-n+1} \ldots w_{i-1})$ en los datos W, con una secuencia de palabras suficientemente largo, se puede aproximar como [2]:

$$H(W) = -\frac{1}{N_W}\log_2 P(W) \quad (12)$$

Donde N_w es la longitud del texto *W* medido en palabras.

La perplejidad *PP(W)* de un modelo de lenguaje *P(W)* se define como el recíproco de la probabilidad media (geométrica) asignado por el modelo a cada palabra en el conjunto de prueba W. Esta es una medida, relacionada con la entropía cruzada, conocida como conjunto de prueba de perplejidad:

$$PP(W) = 2^{H(W)} \quad (13)$$

La perplejidad puede interpretarse aproximadamente como la media geométrica del factor de ramificación del texto cuando se presenta al modelo de lenguaje. La perplejidad definida en la ecuación (**13**) tiene dos parámetros clave: un modelo de lenguaje y una secuencia de palabras. En general, es cierto que una perplejidad menor se correlaciona con un mejor rendimiento del reconocimiento. Un lenguaje con alta perplejidad significa que el número de palabras de ramificación a partir de una palabra previa es alto en promedio. En este sentido, la perplejidad es una indicación de la complejidad del lenguaje si tenemos una estimación exacta de *P(W)* [2].

Los N-gramas son esenciales en cualquier tarea en la que se tengan que identificar palabras en ambientes ruidosos o entradas ambiguas. Por ejemplo, en el reconocimiento del habla, los sonidos de voz de entrada se confunden fácilmente y muchas palabras suenan extremadamente similares, entonces al dar una intuición de que secuencia de palabras van a ingresar, se le ayuda así al reconocedor [1]. En la práctica, se realiza entonces una selección de palabras a fin de minimizar la proporción del vocabulario al sistema mediante la inclusión de las palabras a reconocer.

En el presente proyecto se trabaja con un vocabulario cerrado, en el que se conoce todas las palabras que pueden ocurrir, y por lo tanto, se sabe el tamaño del vocabulario *V* con antelación. En un vocabulario cerrado se parte de que tenemos un léxico, y el conjunto de prueba sólo puede contener palabras de este léxico. Así pues, la tarea del vocabulario cerrado asume que no existen palabras desconocidas [1], en consecuencia se reduce en gran medida la perplejidad.

2.4 Decodificador

El proceso de decodificación en el funcionamiento de un reconocedor de voz es encontrar una secuencia de palabras cuyo correspondiente modelo acústico y modelo de lenguaje mejor se adapte a la secuencia de vector de características de entrada. Por lo tanto, el proceso de decodificación se refiere a menudo como un proceso de búsqueda. La complejidad del algoritmo de búsqueda está altamente correlacionada con el espacio de búsqueda, que está determinada por las restricciones impuestas por los modelos de lenguaje [2].

El decodificador, por lo tanto, descubre la secuencia de palabras $\widehat{W} = w_1 w_2 ... w_m$ que tienen la máxima probabilidad posterior $P(W|X)$ para la observación acústica dada e $X = X_1 X_2 \ldots X_n$, De acuerdo con la operación de maximización descrita en la ecuación (**6**) [2].

Partiendo de que el modelo de lenguaje se realice con HMM, se puede referenciar que la mayoría de los sistemas de reconocimiento de voz utilizan el algoritmo de Viterbi para la decodificación [48] [1] [2]. Las razones para elegir el decodificador de Viterbi implican argumentos que apuntan a que el habla es un proceso de izquierda a derecha y a las eficiencias que ofrece en un proceso de tiempo sincrónico [2].

El algoritmo de Viterbi parte de un estado inicial y, teniendo en cuenta la probabilidad de transición entre estados, la probabilidad de emisión de estos estados y las probabilidades que gobiernan la concatenación de modelos representativos de palabras, obtiene de manera recurrente la secuencia de estados más probable. Los modelos que subyacen a esta secuencia de estados más probable son los que determinan la trascripción de la secuencia que estamos reconociendo [54].

Cada celda de la malla de Viterbi, $v_t(j)$, representa la probabilidad de que el HMM esté en el estado j y haber observado los t primeros vectores de parámetros que pasa a través de la secuencia de estados más probable $q_1 ... q_{t-1}$, dado el autómata λ. El valor de cada celda $v_t(j)$ se calcula tomando recursivamente el camino más probable que nos podría llevar a esta celda. Formalmente, cada celda expresa la siguiente probabilidad [1]:

$$v_t(j) = P(q_0, q_1 .. q_{t-1}, O_1, O_2 \ldots O_t, q_t = j|\lambda) \qquad (14)$$

Una vez se ha calculado la probabilidad de estar en todos los estados en el tiempo $t-1$, se calcula la probabilidad de Viterbi tomando la extensión del camino más probable que conducen a la celda actual. Para un determinado estado q_j en el tiempo t, el valor $v_t(j)$ es calculado como:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(O_t) \quad (15)$$

Donde, $v_{t-1}(i)$ representa la probabilidad de ruta anterior de Viterbi desde el paso de tiempo anterior; a_{ij} representa la probabilidad de transición desde el estado previo q_i hasta el estado actual q_j ; y $b_j(O_t)$ representa la probabilidad de la observación de estado del símbolo de observación O_t dado el estado actual j [1].

Este algoritmo se puede visualizar por medio de una rejilla como se mostrada en la Figura 9. El eje vertical representa a los estados y el eje horizontal el trascurso del tiempo. Esa rejilla muestra todos los caminos posibles que enlazan el nodo inicial con el final. Cuando dos caminos confluyen al mismo nodo únicamente sobrevive el camino que hasta ese punto era el más probable [54].

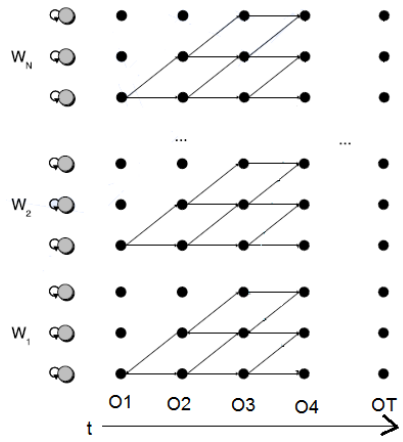

Figura 9: Esquemático de un enrejado de Viterbi para un modelo de lenguaje. Adaptado de [1]

2.5 Problemas en la detección

Uno de los grandes retos a los que los sistemas de reconocimiento de voz se deben enfrentar se relaciona con la gran cantidad de variables presentes en la señal de entrada. Una de ellas se asocia con las características del hablante (como lo son el estilo, tono y ritmo del habla, la fisiología, género, edad y acento del hablante) [47]; es imposible que un locutor (y con más razón varios locutores) pronuncie dos veces exactamente igual una misma sílaba, palabra o frase. Si una persona susurra, grita o canta para reflejar sus estados emocionales, la entonación cambia significativamente. La velocidad del habla también afecta en el reconocimiento, por lo general, cuanto mayor sea la tasa de habla (palabras / minuto), mayor es la tasa de error; Además los patrones del habla de una persona pueden ser totalmente diferentes a los de otra, ya que éstos dependen del tamaño físico de su tracto

vocal, la longitud y anchura del cuello que dependen en gran medida de la edad y el sexo y dan lugar a variaciones en la escala de frecuencias. También son importantes el estado de salud y su condición física (cansancio, gripa, etc.), la educación, su estilo personal, la procedencia geográfica, entre otros [50] [47]. En particular, quienes hablan con algún acento tienen un notable incremento en la tasa de error de 2 a 3 veces [47].

Otras condiciones adversas importantes la constituyen el entorno y el canal de transmisión. El ruido de ambiente acústico suele considerarse aditivo y es la más importante de las posibles condiciones adversas con que el reconocedor puede enfrentarse, además de que las fuentes de este tipo de ruido son abundantes [50]. También debe considerarse que el ruido puede estar presente desde el mismo dispositivo de entrada, como lo es el micrófono y ruidos de interferencia A/D (análogo a digital), así como características frecuenciales de la línea de transmisión. El tipo y ubicación del micrófono, puede añadir así mismo ruido y distorsionar significativamente el espectro de la señal [50] [43]. También afecta las interferencias y reverberaciones de la propia sala. Además, también han de tenerse en cuenta las variaciones en el modo de articular del hablante debido a su reacción psicológica al entorno ruidoso.

Los cambios articulatorios debido a la influencia del entorno, conocidos como efecto Lombard, pueden tener efectos dramáticos en los resultados de reconocimiento. Así, por ejemplo, se ha observado que cuando un locutor habla en presencia de ruido el primer formante de una vocal tiende a crecer mientras que el segundo decrece y que la caída espectral decrece en las frecuencias bajas y aumenta en las altas para la mayoría de las vocales [50].

Por lo descrito anteriormente, los sistemas de reconocimiento de voz se enfrentan a un gran reto, ya que son muchos los problemas que constituyen las causas de degradación de éstos sistemas cuando se usan en la práctica.

3. DESARROLLO METODOLÓGICO

3.1 Análisis de las principales plataformas en software para implementar sistemas de reconocimiento de voz.

Como se referenció en la sección de antecedentes del presente proyecto, a nivel de Software, se encuentran diversos toolkits (conjunto de herramientas) para reconocimiento de voz, dentro de estos se destacan “Julius” desarrollado por diferentes Instituciones de Japón, “CMU Sphinx” desarrollado por la Universidad de Carnegie Mellon – EEUU, “ISIP” desarrollado por la Universidad del estado de Mississippi, “Microsoft Speech API” desarrollado por la compañía Microsoft y especializado en HMM está “HTK” desarrollado por Cambridge University.

Adicionalmente, como apoyo a las anteriores herramientas se encuentra el proyecto Voxforge [38], el cual tiene como objetivo el recoger transcripciones de textos mediante voz que la gente done para ser usada con herramientas de reconocimiento de voz libre y de Código Abierto. Todas las voces que se recogen están disponibles bajo la licencia GPL (Licencia Pública General de GNU), y se usan para 'compilar' modelos acústicos en diversos idiomas para ser usado con software Open Source de reconocimiento de voz. Actualmente VoxForge tiene como objetivo el reconocimiento de voz no orientado al dictado, pero en un futuro, cuando la cantidad de grabaciones de las que se dispongan aumenten, los datos podrían ser usados para la creación de modelos acústicos para software de dictado (para la primera aplicación se estima que se necesitan unas 140 horas de audio, mientras que para sistemas de dictado unas 2000 horas) [38]. Los idiomas que ya disponen de modelos acústicos descargables a través de Voxforge son el alemán, el inglés, el holandés y el ruso. Pese al esfuerzo que está haciendo este proyecto, aún no se dispone de un modelo acústico en español para poder ser descargado, debido a que se necesitan muchas más grabaciones de voz, en consulta realizada en Noviembre del 2014, la página del proyecto muestra en sus estadísticas que se ha logrado recolectar 47.9 horas de grabación de las 140 necesarias para crear el modelo acústico.

3.1.1 HTK (Hidden Markov Model Toolkit)

HTK [36] es un toolkit que ha sido ampliamente utilizado en investigaciones de reconocimiento de voz y se emplea para la construcción y manipulación de modelos ocultos de Markov (HMM), los cuales como se mencionó en el marco teórico constituyen la herramienta matemática más efectiva actualmente para implementar sistemas de reconocimiento del habla. HTK está conformado por un

conjunto de módulos de librerías y herramientas disponibles en forma de archivos de código fuente escritas en lenguaje C. Las herramientas proporcionan facilidad para trabajar con el análisis del habla, entrenamiento de HMM, pruebas y análisis de resultados.

La primera versión fue desarrollada por “Speech Vision and Robotics Group” del departamento de ingeniería de la Universidad de Cambridge (CUED) en 1989 por el profesor Steve Young. En 1993 el Entropic Research Laboratory Inc. adquirió los derechos para vender HTK y el desarrollo de HTK fue totalmente transferido a Entropic en 1995 cuando se estableció el Entropic Cambridge Research Laboratory Ltd. En 1999 Microsoft compró Entropic. Luego Microsoft licenció HTK de nuevo a CUED para que pueda redistribuir y continuar con el desarrollo del software, llamado como HTK3, así como proporcionar soporte para el desarrollo a través del sitio web HTK3 [36].

HTK está disponible para su descarga gratuita pero primero debe estar de acuerdo con la licencia que se presenta, la misma permite usar HTK3 para la enseñanza y la investigación académica, así como utilizarlo para entrenar modelos que luego serán utilizados en productos comerciales. Se puede usar para construir un producto, pero no se permite redistribuir (partes de) htk3 y se debe referencia el uso de HTK en cualquier publicación que haga uso de éste Software.

En la página oficial de HTK se indica que la versión distribuida de HTK3 corre en Linux, Solaris, IRIX, HPUX, Mac OS/Xand FreeBSD. También ha corrido en Windows 2000 y XP, pero no se hace referencia a versiones posteriores de Windows.

3.1.2 Julius

Julius [27] es un software decodificador de código abierto para el reconocimiento continuo del habla en grandes vocabularios, enfocado a desarrolladores e investigadores relacionados con este tema, y que busca promover los avances recientes en estudios de reconocimiento de voz de alto estándar entorno a la comunidad abierta.

Éste software realiza la decodificación con poco gasto de memoria. El núcleo de la aplicación está implementada como una librería embebida, con el propósito de ofrecer la capacidad de reconocimiento de voz a varias aplicaciones; además soporta reconocimiento de entrada de voz en vivo a través de un dispositivo de captura de audio. Julius posee una arquitectura modularizada con el fin de ser independiente de estructuras de modelos, así puede adoptar formatos estándar de otras herramientas libres para modelado como HTK. Esto toma importancia una vez que la distribución de

Julius sólo incluye modelos acústicos completos para el Japonés, por lo que al entrenar otros modelos usando por ejemplo la herramienta HTK podría funcionar en otros idiomas. Julius también permite importar los modelos acústicos ya generados en otros idiomas por la organización VoxForge.

Julius ha sido desarrollado como un Software de investigación para los sistemas de reconocimiento de habla en japonés desde 1997. En su desarrollo han intervenido el Kawahara Lab. de Kyoto University, el Shikano Lab. de Nara Institute of Science and Technology y el equipo del proyecto Julius de Nagoya Institute of Technology. Fue fundado por el Proyecto del Programa de Tecnologías de la Información Avanzada de “Information-technology Promotion Agency” (IPA). En el año 2000 pasó al Consorcio de reconocimiento de voz continua de Japón (CSRC) y desde el 2003 pasó al Consorcio Interactivo de Tecnología del Habla (ISTC). La última versión del programa fué liberada en enero 15 de 2014, bajo la versión 4.3.1

Éste software se desarrolla bajo Linux y Windows. También puede correr en Solaris, FreeBSD and MacOS X. Dado que Julius está escrito en C puro y tiene poca dependencia de librerías externas, puede funcionar en otras plataformas. La página oficial indica que desarrolladores han portado Julius a celulares con Windows, iPhone y otros ambientes con microprocesador.

Julius para SAPI es la versión que se implementa para Microsoft Windows en su Speech API (SAPI) 5.1, pero se debe tener en cuenta que Julius asume que el idioma del usuario y la gramática de la aplicación son en japonés, por lo que para otros lenguajes, Julius para SAPI no conoce en la gramática la pronunciación de las palabras [27].

Es distribuido con licencia abierta, junto con los códigos fuente. Se puede utilizar para cualquier propósito, incluyendo los comerciales y se debe referenciar el uso de Julius en cualquier publicación que haga uso de éste Software.

3.1.3 CMU Sphinx

El toolkit CMU Sphinx [31] desarrollado por la Universidad de Carnegie Mellon – EEUU, describe un grupo de sistemas de reconocimiento de voz usado para construir aplicaciones del habla. Éste es un sistema de reconocimiento de voz continua, de gran vocabulario e independiente del locutor.

Cada uno de los paquetes del CMU Sphinx tiene diferentes tareas y aplicaciones que se listan a continuación [31]:

- Pocketsphinx — librerías para reconocimiento de voz ligeras escritas en C que puede ser utilizado en sistemas embebidos.
- Sphinxbase — librerías de soporte requeridas por Pocketsphinx.
- CMUclmtk — language model tools. Una suite de herramientas que llevan a cabo entrenamiento del modelo del lenguaje.
- Sphinxtrain —Conjunto de herramientas que llevan a cabo el entrenamiento del modelo acústico.
- Sphinx3 — decodificador para la investigación de reconocimiento de voz escrito en C.
- Sphinx4 — Una versión completa de Sphinx escrita en Java que proporciona alta precisión.

Respecto a las plataformas bajo las que corre, CMU Sphinx por su naturaleza de código abierto recomiendan Linux u otras estaciones de trabajo Unix, sin embargo las versiones más modernas también corren bajo windows. Dentro de los lenguajes de programación en los que se puede trabajar esta C, Perl, Java y Python. CMU Sphinx se distribuye bajo la licencia de software libre permisiva BSD, teniendo en cuenta hacer referencia del uso de CMU Sphinx en cualquier publicación que haga uso de este Software.

Esta aplicación, actualmente tiene a disposición un modelo acústico para el español, el cual fue entrenado con datos acústicos que soportan tanto el reconocimiento de banda ancha de las grabaciones de voz del micrófono como el reconocimiento de banda estrecha en el habla por teléfono. El primer y único reporte en su página oficial hasta el momento de una aplicación que utilice dicho modelo acústico, tiene fecha de Noviembre 4 de 2014. La aplicación reportada controla las acciones de un sillón con las palabras “Activación Sillón”, “Sube”, “Baja” y “Parar”, logrando buenos resultados.

3.1.4 Microsoft SAPI

Un avance importante en el desarrollo en tecnologías del habla de Microsoft se dio en 1993, cuando como se indica en [56], ésta compañía “contrató a Xuedong (XD) Huang, Fil Alleva, y Mei-Yuh Hwang, tres de las cuatro personas responsables del sistema de reconocimiento de voz Sphinx -II de la Universidad de Carnegie Mellon. Desde el principio, con la formación del equipo de la API de voz (SAPI - Interfaz de Programación de Aplicaciones de Voz) 1.0 en 1994, Microsoft fue impulsado a

crear una tecnología de voz que fuera a la vez precisa y accesible a desarrolladores a través de un potente API", lanzando así en los últimos años una serie de plataformas cada vez más potentes para el desarrollo del reconocimiento y síntesis de voz en su sistema operativo Windows y disponible para una gran variedad de idiomas [56].

Particularmente en Windows 7, sistema operativo en el que se desarrolló la aplicación del presente proyecto, hay dos APIs del habla [56]:

- SAPI 5.4, el API de código nativo para la programación de los motores de voz incluidos en Windows 7.
- Los espacios de nombres System.Speech, para la implementación del habla en aplicaciones de Windows usando código gestionado como por ejemplo C # en Microsoft .NET Framework. Estos espacios de nombres están disponibles en Microsoft .NET Framework 3.0 y versiones posteriores e incluye los paquetes System.Speech.Synthesis y System.Speech.Recognition para la conversión de texto a voz y para el reconocimiento del habla respectivamente.

Ambos tipos implementan el SAPI DDI (interfaz de controlador de dispositivo), que es una API que hace a los motores intercambiables para las capas por encima de ellos, lo que significa que los desarrolladores que utilizan SAPI o System.Speech son libres de utilizar otros motores que implementan el SAPI DDI [56].

SAPI reduce drásticamente la sobrecarga de código requerido para que una aplicación utilice el reconocimiento de voz y la conversión de texto a voz, haciendo que estas tecnologías sean más accesibles y robustas para un amplio rango de aplicaciones. SAPI proporciona así una interfaz de alto nivel entre una aplicación y los motores de voz, implementando todos los detalles de bajo nivel necesarios para controlar y gestionar las operaciones en tiempo real de sus motores de voz [37].

Los dos tipos básicos de motores SAPI son los sistemas de conversión de texto a voz (TTS - text-to-speech) y los sistemas de reconocimiento de voz. Los primeros sintetizan las cadenas de texto y archivos en audio hablado usando voces sintéticas; por su parte, los sistemas de reconocimiento de voz convierten el audio del habla humana en archivos y cadenas de texto legibles [37]. SAPI reconoce sin problema el idioma Español siempre y cuando el sistema operativo en el equipo que lo esté corriendo tenga instalado dicho idioma y se va entrenando conforme se va haciendo uso del mismo.

Al usar el espacio de nombres System.Speech, se le puede dar a las aplicaciones de Windows la capacidad de responder a comandos de voz, aceptar dictado, generar voz sintetizada (conversión de texto a voz), o reproducir archivos de audio pregrabados. Específicamente con el espacio de nombres System.Speech.Recognition, se proporciona la funcionalidad para adquirir y monitorear la entrada de voz, crear gramáticas de reconocimiento del habla que produzcan tanto resultados de reconocimiento literales como semánticos, capturar información de eventos generados por el reconocedor de voz y configurar y administrar los motores de reconocimiento del habla [57]. System.Speech.Recognition soporta la especificación W3C para la Gramática de Reconocimiento de habla, documentada en [58] . Pero para la mayoría de los casos este método es exagerado, por lo que la API también proporciona la clase GrammarBuilder que permite construir una gramática de un conjunto de frases y opciones [56].

Referente a las técnicas que utiliza Microsoft para el modelo acústico, se referencia en [59] el uso de un híbrido entre un pre entrenamiento de redes neuronales profundas (DNN) y un modelo oculto de Markov (HMM) dependiente del contexto (CD) para el reconocimiento del habla en vocabulario largo, técnica reconocida con la abreviación CD-DNN-HMMs. Esta arquitectura híbrida entrena las redes neuronales profundas para producir una distribución sobre senones (estados de trifonemas atados) como sus salidas. El entrenamiento de redes neuronales para predecir una distribución sobre senones brinda mayor cantidad de bits de información que estarán presentes en las etiquetas de la red neuronal entrenada [59].

En cuanto a la licencia, .NET Framework, que contiene el System.Speech, es a la fecha un complemento de código cerrado gratuito para los sistemas operativos Windows licenciados, disponible para todos los usuarios, incluidos los desarrolladores que crean aplicaciones .NET quienes son libres de redistribuir .NET Frameworks con su aplicación.

Cabe destacar que Microsoft anunció en su centro de noticias oficial, el 13 de Noviembre de 2014 que dicha compañía ofrecerá nuevas funcionalidades y mejoras para los desarrolladores y toda la comunidad open source en Visual Studio 2015, .NET 2015 y Visual Studio Online. Dentro del comunicado se resalta: “Así, fiel a su compromiso de apoyar el desarrollo multiplataforma, Microsoft proporcionará el rango completo .NET server stack en código abierto, incluyendo ASP.NET, el compilador .NET, .NET Core Runtime, Framework y Librerías, de forma que permitirá a los desarrolladores trabajar con .NET en todas las plataformas, Windows, Mac OS o Linux. Se trata de un paso sin precedentes en la estrategia de Microsoft, que trabajará estrechamente a través de .NET

Foundation con la comunidad de código abierto en pro de lograr grandes avances y mejoras en el entorno .NET". También se anunció la versión Preview de Visual Studio 2015 que como se indica en el mismo comunicado, "con soporte para iOS, Android y Windows, facilita a los desarrolladores construir aplicaciones y servicios para cualquier dispositivo y en cualquier plataforma" [60].

3.2 Plataforma en software Seleccionada

Para el proyecto "reconocimiento de comandos de voz en español orientado al control de una silla de ruedas" se desarrolló la aplicación de reconocimiento de voz utilizando Microsoft SAPI en un entorno de escritorio de Windows 7, usando los espacios de nombres System.Speech y especialmente System.Speech.Recognition en el Microsoft .NET Framework 4.

La principal razón para seleccionar dicha plataforma se debe al modelo acústico que para el español ya poseía Microsoft .NET Framework, el cual permite un reconocimiento independiente del hablante y sin necesidad de mayores entrenamientos previos, además de que el comportamiento de su motor de voz ya era conocido, al utilizar SAPI 5.4 incluido en Windows 7 para interactuar por comandos de voz con dicho sistema operativo y sus aplicaciones. HTK por su parte fue descartado puesto que el objetivo del proyecto no es construir un modelo acústico, si no por el contrario partir de un modelo ya existente en español y lo suficientemente completo como para realizar el reconocimiento independientemente del hablante y entrenado con un vocabulario tan amplio que no se tenga problemas al momento de incluir nuevas secuencias de palabras en el vocabulario a reconocer. Julius se descartó ya que solo dispone de un modelo acústico completo para el Japonés y CMU Sphinx no tenía reportes al momento de comenzar el desarrollo del presente proyecto sobre el comportamiento del modelo acústico que estaba en construcción en idioma español, la primera aplicación reportada solo aparece con fecha de Noviembre 4 de 2014.

Tampoco se tiene problema con lo referente al licenciamiento de Windows y por consecuente al uso de .Net Framework, ya que la Universidad Autónoma de Manizales cuenta con equipos licenciados con este sistema operativo, al igual que el computador personal en el que se desarrolló la aplicación. El lenguaje de programación utilizado fue C#, desarrollado en Visual Studio 2008.

3.3 Arquitectura del sistema

La herramienta utilizada para documentar los requerimientos funcionales y no funcionales, así como los casos de uso detallados fue REM (REquirement Manager). El diagrama de casos de uso se realizó en StarUML y los diagramas de objetos, de clases, de secuencia y de despliegue se realizaron utilizando la herramienta Modelio 3.1.

3.3.1. Requerimientos Significativos de la arquitectura

En esta subsección se identifican los requerimientos funcionales y los requerimientos no funcionales del sistema.

3.3.1.1 Requerimientos Funcionales

FRQ-0001	Reconocer comandos de voz relacionados con el control del desplazamiento de la silla
Versión	1.0 (12/02/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *reconocer palabras asociadas al control de movimiento de la silla de ruedas. Tales como izquierda, derecha, adelante, atrás, parar, lento y rápido,*
Importancia	vital
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0003	Reconocer comandos de voz orientado a ordenes de Domótica
Versión	1.0 (05/03/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *reconocer frases asociadas a ordenes de domótica. Tales como "prender luz entrada", "cerrar cortina alcoba", "apagar luz pasillo", entre otros.*
Importancia	vital
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0004	Reconocer comandos de voz asociados a medición de signos vitales.
Versión	1.0 (05/03/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *reconocer palabras asociadas con la adquisición de signos vitales. Tales como "capturar electro", "medir temperatura", "tomar presión", "medir pulso", entre otras.*
Importancia	vital
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0002	Gestionar el envío de correos electrónicos a destinatarios predeterminados por el usuario
Versión	1.0 (12/02/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *permitir la gestión del envío de correos electrónicos por comandos de voz.* *Este requirimiento incluye el asociar comandos de voz con direcciones de correo electronico, personalizar las plantillas de envío para cada destinatario, y observar y eliminar direcciones de correo ya almacenadas.*
Importancia	vital
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0005	configurar parámetros del sistema de reconocimiento
Versión	1.0 (05/03/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *dar la opción para configurar algunos parametros asociados con el sistema de reconocimiento.*
Importancia	vital
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0006	Gestionar la apertura de aplicaciones
Versión	1.0 (05/03/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *permitir la gestión de lanzar aplicaciones instaladas en el computador del usuario.*
Importancia	importante
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0007	Administar plantilla de remitente de correo electrónico
Versión	1.0 (05/03/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *administrar la plantilla con los datos de la dirección de correo electrónico del remitente (usuario de la silla). Esta misma plantilla tendrá el asunto y mensaje predeterminado para todo nuevo destinatario.*
Importancia	vital
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0008	visualizar variables del estado y resultado del reconocimiento
Versión	1.0 (05/03/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *permitir la visualización de variables referentes al estado y resultado del reconocimiento de los comandos. Datos necesarios para las pruebas del sistema.*
Importancia	vital
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0009	Visualizar el comando reconocido
Versión	1.0 (26/11/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *mostrar el comando reconocido en caso de ser exitoso, así como indicar su no identificación en caso de no ser reconocido.*
Importancia	importante
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0010	Visualizar las palabras a pronunciar para el reconocimiento de los comandos de voz
Versión	1.0 (26/11/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *mostrar las palabras a pronunciar para el reconocimiento de los comandos de voz referentes al control del movimiento de la silla, a las órdenes de domótica y a las de la toma de variables del cuerpo humano.*
Importancia	importante
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

FRQ-0011	visualizar la plantilla de envio de correo
Versión	1.0 (30/11/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *mostrar la plantilla a enviar según destinatario seleccionado, ésta se visualizará en el momento en que el usuario pronuncia el comando que selecciona a uno de los destinatarios.*
Importancia	importante
Urgencia	hay presión
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

3.3.1.2 Requerimientos No Funcionales

NFR-0001	Indicador de entrada de audio
Versión	1.0 (12/02/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Lily Jhohana Gil Vásquez
Dependencias	Ninguno
Descripción	El sistema deberá *mostrar en una barra de progreso la intensidad de la señal de entrada de audio.*
Importancia	quedaría bien
Urgencia	puede esperar
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

NFR-0005	Proponer posibles comandos a pronunciar
Versión	1.0 (12/03/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *proponer posibles comandos que quizas el usuraio desea decir cuando el sistema detecta un comando pronunciado, pero el nivel de confidencia no alcanza el mínimo establecido para ser aceptado como válido.*
Importancia	quedaría bien
Urgencia	inmediatamente
Estado	pendiente de validación
Estabilidad	alta
Comentarios	Ninguno

NFR-0006	Tiempo de respuesta
Versión	1.0 (01/12/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *responder al reconocimiento de los comandos de voz tan pronto el usuario pronuncie algún comando.*
Importancia	importante
Urgencia	inmediatamente
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

NFR-0007	Entorno de Desarrollo
Versión	1.0 (01/12/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *correr bajo el sistema operativo Windows, ya que la plataforma de desarrollo a utilizar será .Net Framework*
Importancia	importante
Urgencia	hay presión
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

NFR-0008	Interfaz de usuario
Versión	1.0 (01/12/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *tener una interfaz de usuario basada en formularios y pestañas que faciliten la visualización de las diferentes funciones del aplicativo.*
Importancia	importante
Urgencia	hay presión
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

NFR-0009	Disponibilidad
Versión	1.0 (01/12/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *lanzarse automáticamente una vez el usuario prende el computador.*
Importancia	importante
Urgencia	hay presión
Estado	validado
Estabilidad	alta
Comentarios	Ninguno

NFR-0010	Interfaces de conexión
Versión	1.0 (01/12/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Dependencias	Ninguno
Descripción	El sistema deberá *ser desarrollado en un lenguaje de programación que permita facilmente la interconexión con el puerto USB para una futura integración del software con la plataforma de hardware Arduino.*
Importancia	importante
Urgencia	puede esperar
Estado	pendiente de validación
Estabilidad	alta
Comentarios	Ninguno

3.3.2. Casos de uso

3.3.2.1 Actor

ACT-0001	Paciente o asistente del paciente
Versión	1.0 (12/02/2014)
Autores	• Lily Jhohana Gil Vásquez
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez
Descripción	Este actor representa *al usuario final (paciente), pero también puede estar representado por el asistente del paciente en caso de que éste último no pueda configurar por si mismo las opciones del programa.*
Comentarios	Ninguno

3.3.2.2 Diagrama de casos de uso

Este diagrama captura información sobre los servicios que el sistema proporciona a su entorno, desde el punto de vista del usuario.

Figura 10: Diagrama de casos de uso

3.3.2.3 Casos de uso detallado

UC-0001	**Configurar parámetros de reconocimiento**	
Versión	1.0 (12/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea establecer los valores de confidencia mínimo y/o espera de tiempo final.*	
Precondición	Ninguna	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *ingresa los valores que desea establecer para los parámetros de confidencia mínima y el tiempo de espera final.*
	2	El sistema *registra los valores ingresados por el usuario.*
Postcondición	El valor de confidencia mínimo y/o espera de tiempo final establecidos.	
Excepciones	**Paso**	**Acción**
	-	-
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0005	**Grabar parámetros de reconocimiento**	
Versión	1.0 (23/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea conservar los valores establecidos para confidencia mínimo y/o tiempo de espera final una vez cierre la aplicación.*	
Precondición	Valores de confidencia mínimo y/o tiempo de espera final	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *da clic en el botón "Grabar configuración reconocimiento".*
	2	El sistema *actualiza el archivo de configuración de parámetros de audio con los valores de confidencia mínimo y tiempo de espera final dados por el usuario.*
	3	El sistema *despliega un mensaje informando que la configuración del audio fué grabada con éxito*
Postcondición	El valor de confidencia mínimo y/o espera de tiempo final se almacena.	
Excepciones	**Paso**	**Acción**
	2	Si *previamente no hay archivo de configuración con parámetros de audio,* el sistema *crea éste archivo en la ruta de configuración del programa,* a continuación este caso de uso *continúa*
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0006	**Configurar cuentas de correo destinatario**	
Versión	1.0 (23/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea asociar un comando de voz con una cuenta de correo de algún destinatario.*	
Precondición	Ninguna	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *digita la palabra que asociará como comando de voz a un correo de destinatario específico.*
	2	El actor Paciente o asistente del paciente (ACT-0001) *digita la dirección de correo electrónico del remitente que desea asociar con el comando de voz ingresado en el paso 1.*
	3	El actor Paciente o asistente del paciente (ACT-0001) *da clic en el botón que agregará los datos digitados como comando y dirección de correo a la lista de correos de destinatario*
	4	El sistema *registra los valores ingresados por el usuario, actualizando la lista de correos de destinatario así como el archivo de configuración de correos*
Postcondición	Cuantas de correo asociadas a un comando.	
Excepciones	**Paso**	**Acción**
	4	Si *el campo para digitar el comado de voz esta vacio,* el sistema *sacará un mensaje pidiendo que digite el comando de voz para el correo,* a continuación este caso de uso *queda sin efecto*
	4	Si *el campo para digitar la dirección de correo electrónico esta vacio,* el sistema *sacará un mensaje pidiendo que digite el correo,* a continuación este caso de uso *queda sin efecto*
	4	Si *previamente no hay archivo de configuración de correos,* el sistema *crea éste archivo en la ruta de configuración del programa,* a continuación este caso de uso *continúa*
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0012	**Eliminar cuentas de correo destinatario**	
Versión	1.0 (24/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea eliminar un correo electrónico y su comando asociado de la lista de destinatarios*.	
Precondición	Por lo menos debe existir en la lista de correos una cuenta de correo de remitente .	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *da clic en el botón de eliminar que corresponde al correo de destinatario que desea eliminar en la lista de correos*
	2	El sistema *confirma mediante una pregunta si se desea eliminar dicha dirección de correo de la lista*
	3	El actor Paciente o asistente del paciente (ACT-0001) *confirma o rechaza el deseo de eliminar el correo*
	4	Si *el usuario confirma su intensión,* el sistema *elimina el correo de la lista de destinatarios*
	5	Si *el usuario rechaza la intensión,* el sistema *no realiza cambios en la lista de correos de destinatarios.*
Postcondición	Se elimina de la lista de correos los campos pertenecientes al destinatario deseado.	
Excepciones	**Paso**	**Acción**
	-	-
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	importante	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0003	**Configurar plantilla de correo**	
Versión	1.0 (12/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea configurar la plantilla de destinatario o de remitente*	
Precondición	Ninguna	
Secuencia normal	**Paso**	**Acción**
	1	Si *el paciente o asistente del paciente da clic en el botón de la plantilla general de correo,* el sistema *mostrará la plantilla del remitente*
	2	Se realiza el caso de uso Configurar plantilla remitente (UC-0004)
	3	Si *el paciente o asistente del paciente da clic en el botón de plantilla que se encuentra en la lista de correos para cada dirección de destinatario,* el sistema *mostrará la plantilla del destinatario seleccionado*
	4	Se realiza el caso de uso Configurar plantilla destinatario (UC-0008)
Postcondición	Datos de la plantilla de remitente y/o plantilla de destinatario personalizados.	
Excepciones	**Paso**	**Acción**
	-	-
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0004	**Configurar plantilla remitente**	
Versión	1.0 (12/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea configurar los datos de la cuenta de correo mediante la cual se enviarán los mensajes, así como la plantilla general para los destinatarios en caso de que no se configuren individualmente.* o durante la realización de los siguientes casos de uso: [UC-0003] Configurar plantilla de correo	
Precondición	Ninguna	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *da clic en el botón para configurar la plantilla del remitente*
	2	El sistema *abre la plantilla con los campos a modificar por el usuario. Dichos campos corresponden a la dirección del correo electrónico del remitente y la contraseña de acceso al mismo. Así como a los campos para el nombre que se desea aparezca como remitente de los correos y asunto y texto del mensaje predeterminados para cada destinatario en caso de que no se personalice cada uno de manera independiente.*
	3	El actor Paciente o asistente del paciente (ACT-0001) *completa los campos solicitados y al finalizar da clic en el botón para guardar.*
	4	El sistema *registra los valores ingresados por el usuario*
	5	El sistema *despliega un mensaje informando que la configuración predeterminada de correo fué grabada con éxito.*
Postcondición	Datos de la plantilla de remitente establecidos	
Excepciones	**Paso**	**Acción**
	4	Si *previamente no hay archivo de configuración con los datos de la plantilla general,* el sistema *crea éste archivo en la ruta de configuración del programa*, a continuación este caso de uso *continúa*
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0008	**Configurar plantilla destinatario**		
Versión	1.0 (23/03/2014)		
Autores	• Lily Jhohana Gil Vásquez		
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez		
Dependencias	Ninguno		
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea personalizar el asunto y mensaje predeterminado que se enviará a cada destinatario.* o durante la realización de los siguientes casos de uso: [UC-0003] Configurar plantilla de correo		
Precondición	Destinatarios de correo almacenados		
Secuencia normal	**Paso**	**Acción**	
	1	El actor Paciente o asistente del paciente (ACT-0001) *dentro de la lista de correos da clic en el botón para configurar la plantilla correspondiente del destinatario que desea personalizar.*	
	2	El sistema *abre la plantilla con los campos de asunto y mensaje a modificar correspondientes al destinatario seleccionado en el paso 1*	
	3	El actor Paciente o asistente del paciente (ACT-0001) *completa los campos solicitados y al finalizar da clic en el botón para guardar.*	
	4	El sistema *registra los valores ingresados por el usuario*	
	5	El sistema *despliega un mensaje informando que los datos de la plantilla para el destinatario seleccionado fué grabada con éxito.*	
Postcondición	Datos de la plantilla de cada destinatario establecidos		
Excepciones	**Paso**	**Acción**	
	4	Si *previamente no hay archivo de configuración con los datos de la plantilla personalizada para el destinatario seleccionado,* el sistema *crea éste archivo en la ruta de configuración del programa*, a continuación este caso de uso *continúa*	
Rendimiento	**Paso**	**Tiempo máximo**	
	-	-	
Frecuencia esperada	PD		
Importancia	vital		
Urgencia	inmediatamente		
Estado	validado		
Estabilidad	alta		
Comentarios	Ninguno		

UC-0002	**Enviar correo**		
Versión	1.0 (12/03/2014)		
Autores	• Lily Jhohana Gil Vásquez		
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez		
Dependencias	Ninguno		
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea enviar un correo electrónico.* o durante la realización de los siguientes casos de uso: [UC-0011] Procesar comandos de voz		
Precondición	la lista de correos debe contener la dirección de correo de destinatario al que se le desea enviar el mensaje con su correspondiente comando de voz asociado		
Secuencia normal	**Paso**	**Acción**	
	1	El actor Paciente o asistente del paciente (ACT-0001) *Pronuncia el comando asociado al destinatario de correo.*	
	2	El sistema *carga la plantilla asociada al destinatario*	
	3	El actor Paciente o asistente del paciente (ACT-0001) *Pronuncia el comando "Enviar correo" o da clic en el botón que realizará la misma función*	
	4	El sistema *envía el correo*	
	5	Si *el envío del correo se pudó realizar,* el sistema *despliega un mensaje informando que fué enviado con éxito*	
Postcondición	Correo electrónico enviado		
Excepciones	**Paso**	**Acción**	
	2	Si *no existe plantilla personalizada para el destinatario,* el sistema *carga la plantilla por defecto,* a continuación este caso de uso *continúa*	
	2	Si *el comando asociado al destinatario no se encuentra registrado en la lista de correos,* el sistema *informa que el comando no ha sido reconocido,* a continuación este caso de uso *queda sin efecto*	
	4	Si *no hay conexión a Internet o no se puede obtener respuesta de autenticación éxitosa con la cuenta del remitente,* el sistema *informa que no se pudo enviar el correo,* a continuación este caso de uso *queda sin efecto*	
Rendimiento	**Paso**	**Tiempo máximo**	
	-	-	
Frecuencia esperada	PD		
Importancia	vital		
Urgencia	inmediatamente		
Estado	validado		
Estabilidad	alta		
Comentarios	Ninguno		

UC-0007	**Configurar aplicaciones a ejecutar**	
Versión	1.0 (23/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea asociar un comando de voz con la apertura de alguna aplicación instalada en el computador.*	
Precondición	Ninguna	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *digita la palabra que asociará como comando de voz para la apertura de una aplicación específica.*
	2	El actor Paciente o asistente del paciente (ACT-0001) *digita la ruta de acceso al ejecutable de la aplicación que desea asociar con el comando de voz ingresado en el paso 1.*
	3	El actor Paciente o asistente del paciente (ACT-0001) *da clic en el botón que agregará los datos digitados como comando y ruta del ejecutable a la lista de aplicaciones*
	4	El sistema *registra los valores ingresados por el usuario, actualizando la lista de aplicaciónes así como el archivo de configuración del mismo*
Postcondición	Aplicaciones asociadas a un comando	
Excepciones	**Paso**	**Acción**
	4	Si *el campo para digitar el comado de voz esta vacio,* el sistema *sacará un mensaje pidiendo que digite el comando de voz para la aplicación*, a continuación este caso de uso *queda sin efecto*
	4	Si *el campo para digitar la ruta de la aplicación esta vacio,* el sistema *sacará un mensaje pidiendo que digite la ruta para el programa*, a continuación este caso de uso *queda sin efecto*
	4	Si *previamente no hay archivo de configuración de aplicaciones,* el sistema *crea éste archivo en la ruta de configuración del programa*, a continuación este caso de uso *continúa*
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0013	**Eliminar aplicaciones a ejecutar**	
Versión	1.0 (05/04/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea eliminar de la lista de aplicaciones la apertura por comandos de voz de alguna aplicación*	
Precondición	La lista de aplicaciones debe contener por lo menos la ruta de una aplicación almacenada con su correspondiente comando de voz	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *da clic en el botón de eliminar que corresponde a la aplicación que desea eliminar en la lista de aplicaciones*
	2	El sistema *confirma mediante una pregunta si se desea eliminar dicha aplicación de la lista*
	3	El actor Paciente o asistente del paciente (ACT-0001) *confirma o rechaza el deseo de eliminar la aplicación de la lista*
	4	Si *el usuario confirma su intensión,* el sistema *elimina la aplicación de la lista*
	5	Si *el usuario rechaza la intensión,* el sistema *no realiza cambios en la lista de aplicaciones.*
Postcondición	Se elimina de la lista de aplicaciones los campos pertenecientes a la apertura de la aplicación deseada.	
Excepciones	**Paso**	**Acción**
	-	-
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	importante	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0010	**Detener reconocimiento**	
Versión	1.0 (23/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea que el sistema detenga el reconocimiento de los comandos de voz*. o durante la realización de los siguientes casos de uso: [UC-0011] Procesar comandos de voz	
Precondición	Reconocimiento de voz activado	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *Pronuncia el comando para detener el reconocimiento o da clic en el botón que realizará la misma función*
	2	El sistema *deshabilita las acciones que se realizan por comandos de voz e informa que el reconocimiento ha sido detenido.*
	3	El sistema *deshabilita el botón de detener reconocimiento y habilita el de activar de nuevo el reconocimiento*
Postcondición	Reconocimiento de voz detenido	
Excepciones	**Paso**	**Acción**
	2	Si *el reconocimiento del comando no es exitoso,* el sistema *informa que el comando no ha sido reconocido*, a continuación este caso de uso *queda sin efecto*
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0009	**Activar reconocimiento**	
Versión	1.0 (23/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea que el sistema reconozca de nuevo los comandos de voz pronunciados una vez que el reconocimiento ha sido desactivado.* o durante la realización de los siguientes casos de uso: [UC-0011] Procesar comandos de voz	
Precondición	Reconocimiento de voz detenido	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *Pronuncia el comando para activar de nuevo el reconocimiento o da clic en el botón que realizará la misma función*
	2	El sistema *habilita las acciones que se realizan por comandos de voz e informa que el reconocimiento esta activo.*
	3	El sistema *habilita el botón de detener reconocimiento y deshabilita el de activar de nuevo el reconocimiento*
Postcondición	Reconocimiento de voz activado	
Excepciones	**Paso**	**Acción**
	-	-
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0011	**Procesar comandos de voz**	
Versión	1.0 (23/03/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario pronuncia uno de los comando de voz preestablecidos en la aplicación.*	
Precondición	Reconocimiento de voz activado. Parámatros de reconocimiento configurados. Comandos de voz asociados a direcciones de correo electrónico y a la apertura de aplicaciones instaladas en el computador.	
Secuencia normal	**Paso**	**Acción**
	1	Si *el paciente o asistente del paciente pronuncia el comando para detener el reconocimiento de voz,* se realiza el caso de uso Detener reconocimiento (UC-0010)
	2	Si *el paciente o asistente del paciente pronuncia el comando para activar de nuevo el reconocimiento de voz,* se realiza el caso de uso Activar reconocimiento (UC-0009)
	3	Si *el paciente o asistente del paciente pronuncia el comando para enviar un correo electrónico a un destinatario preconfigurado,* se realiza el caso de uso Enviar correo (UC-0002)
	4	Si *el paciente o asistente del paciente pronuncia el comando para abrir algún programa instalado en el computador,* se realiza el caso de uso lanzar aplicaciones (UC-0014)
	5	Si *el paciente o asistente del paciente pronuncia el comando para ubicarse en una de las pestañas de la aplicación,* se realiza el caso de uso Navegar por las pestañas (UC-0015)
Postcondición	Apertura de aplicaciones instaladas en el computador. Envío de correo electrónico a destinatarios seleccionados. Visualización del comando reconocido. Posicionamiento en la pestaña deseada dentro de la aplicación. Deshabilitación o habilitación del reconocimiento de voz.	
Excepciones	**Paso**	**Acción**
	-	-
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0014	**lanzar aplicaciones**	
Versión	1.0 (06/04/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea abrir alguna de las aplicaciones instaladas en su computador* o durante la realización de los siguientes casos de uso: [UC-0011] Procesar comandos de voz	
Precondición	la lista de aplicaciones debe contener la ruta de acceso al ejecutable de la aplicación con su correspondiente comando de voz asociado. Reconocimiento de voz activado. Parámatros de reconocimiento configurados.	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *Pronuncia el comando de voz asociado a la ruta del ejecutable de la aplicación que desea abrir*
	2	El sistema *abre la aplicación seleccionada en el paso 1.*
Postcondición	aplicación seleccionada abierta	
Excepciones	**Paso**	**Acción**
	2	Si *el comando asociado a la ruta del ejecutable del programa no se encuentra registrado en la lista de aplicaciónes,* el sistema *informa que el comando no ha sido reconocido*, a continuación este caso de uso *queda sin efecto*
	2	Si *la ruta de acceso al ejecutable de la aplicación está erronea,* el sistema *despliega un mensaje informando que se presentó un problema al abrir el programa*, a continuación este caso de uso *queda sin efecto*
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0015	**Navegar por las pestañas**	
Versión	1.0 (06/04/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea visualizar por medio de comandos de voz el contenido de las diferentes pestañas que posee la aplicación.* o durante la realización de los siguientes casos de uso: [UC-0011] Procesar comandos de voz	
Precondición	Reconocimiento de voz activado. Parámatros de reconocimiento configurados.	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *Pronuncia el comando de voz asociado a la pestaña en la cual desea visualizar su contenido*
	2	El sistema *despliega el contenido de la pestaña seleccionada en el paso 1.*
Postcondición	la visualización de la pantalla del programa se situa en la pestaña deseada	
Excepciones	**Paso**	**Acción**
	2	Si *el reconocimiento del comando no es exitoso,* el sistema *informa que el comando no ha sido reconocido*, a continuación este caso de uso *queda sin efecto*
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	importante	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

UC-0016	**Reconocimiento de comandos específicos de la silla**	
Versión	1.0 (06/04/2014)	
Autores	• Lily Jhohana Gil Vásquez	
Fuentes	• Luis Fernando Castillo • Ruben Dario Florez	
Dependencias	Ninguno	
Descripción	El sistema deberá comportarse tal como se describe en el siguiente caso de uso cuando *el usuario desea que se reconozca el comando de voz referente al desplazamiento de la silla, a ordenes de domótica o a la medición de signos vitales*	
Precondición	Reconocimiento de voz activado. Parámatros de reconocimiento configurados.	
Secuencia normal	**Paso**	**Acción**
	1	El actor Paciente o asistente del paciente (ACT-0001) *Pronuncia uno de los comandos de voz asociado al desplazamiento de la silla o a las órdenes de domótica o a la medición de signos vitales que por defecto están configurados en la aplicación*
	2	Si *el comando es reconocido,* el sistema *depliega un texto con la frase pronunciada y mediante una señal auditiva da a conocer el éxito del reconocimiento*
	3	Si *el comando no es reconocido,* el sistema *depliega un texto indicando su estado y pide que intente de nuevo.*
Postcondición	Visualización del comando reconocido	
Excepciones	**Paso**	**Acción**
	-	-
Rendimiento	**Paso**	**Tiempo máximo**
	-	-
Frecuencia esperada	PD	
Importancia	vital	
Urgencia	inmediatamente	
Estado	validado	
Estabilidad	alta	
Comentarios	Ninguno	

3.3.3. Diagrama de objetos

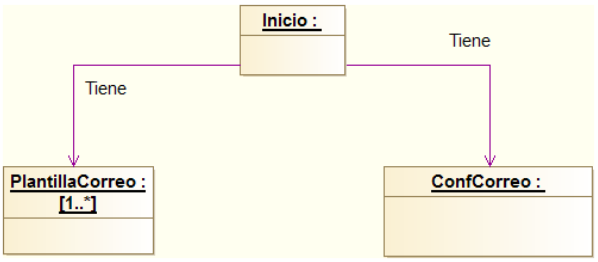

Figura 11: Diagrama de objetos

3.3.4. Diagrama de clases

Se debe aclarar que en todos los métodos que poseen el parámetro llamado “sender”, este es de tipo object y lo acompaña otro parámetro llamado “e” que es de tipo EventArgs. En el diagrama de clases realizado con la herramienta Modelio, no fue posible configurar dichos tipos de dato, por lo que aparecen como si fueran tipo string.

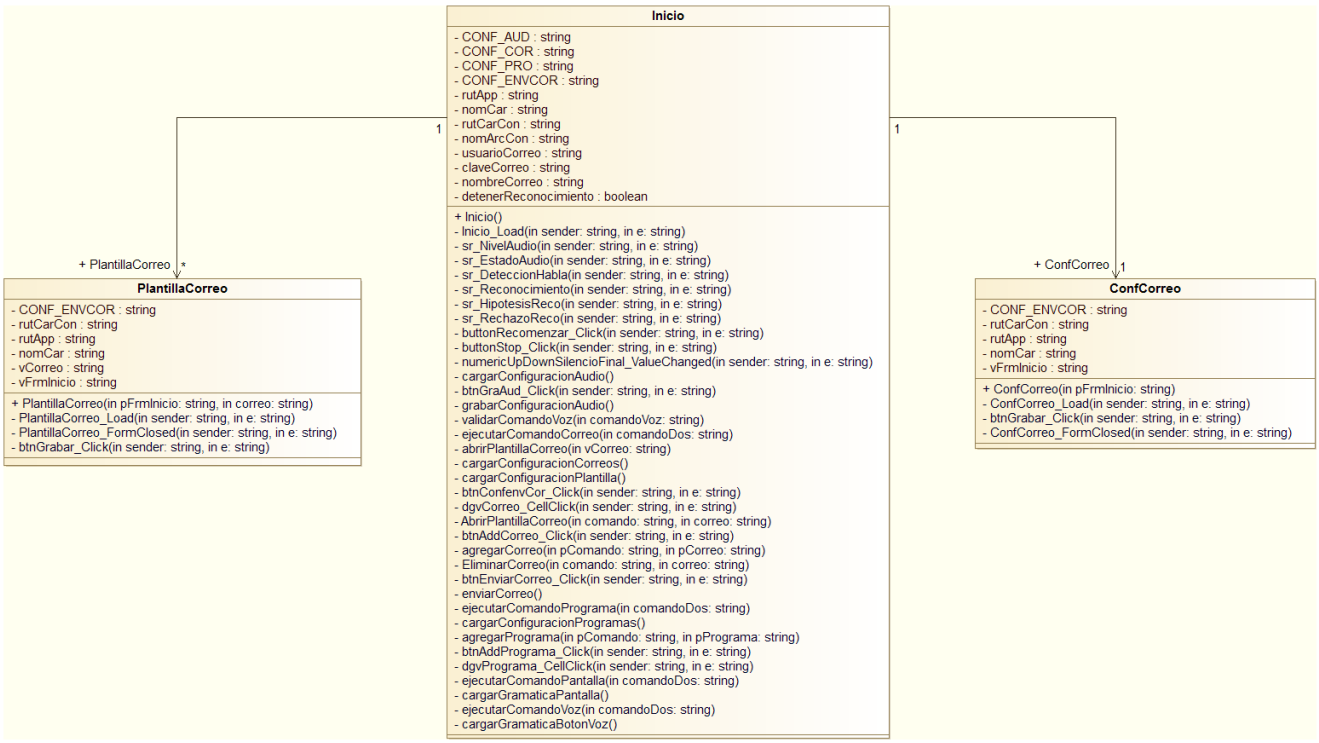

Figura 12: Diagrama de clases

3.3.5. Modelo de despliegue

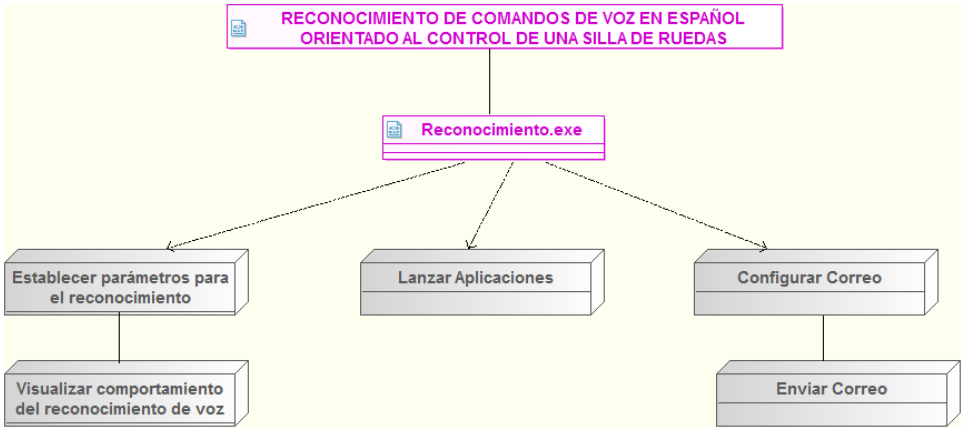

Figura 13: Modelo de despliegue

3.4 Modelo de lenguaje para la aplicación desarrollada bajo un sistema de reconocimiento con vocabulario cerrado

Como ya se ha descrito, el modelo de lenguaje representa la forma en que se combinan las palabras en un idioma.

Después de investigar sobre las diferentes herramientas existentes para el reconocimiento de la voz, descritas en el capítulo 3.1, se encontró que el SAPI de Microsoft tenía ya muy desarrollado un modelo de lenguaje de propósito general para el español, por lo cual se decidió tomar esa herramienta y adaptar mas bien su modelo de lenguaje a las necesidades específicas de la aplicación, en donde solo se requiere reconocer ciertas expresiones de interés para la misma (vocabulario cerrado). Por lo tanto se definió una gramática que limita el reconocedor para escuchar sólo el habla que le interesa a la aplicación.

Al limitar el contenido de la gramática de reconocimiento solo para los comandos disponibles, se obtienen beneficios como [56] [61]:

- Se aumenta la precisión y rendimiento del reconocedor comparado con tareas de dictado (vocabulario abierto). Dicha comparación se hace evidente una vez que un motor de reconocimiento de voz para un vocabulario abierto debe abarcar casi un diccionario entero de la lengua.
- Se garantiza que todos los resultados del reconocimiento tengan significado para la aplicación, y permite al motor de reconocimiento especificar los valores semánticos inherentes en el texto reconocido.
- Reduce la sobrecarga de procesamiento que la aplicación requiere.
- Permite un procesamiento independiente del locutor, lo que elimina la necesidad de entrenar el reconocedor para configurar perfiles por cada hablante.

En el desarrollo de la aplicación, las clases del espacio de nombres System.Speech.Recognition utilizadas en la construcción de la gramática para los comandos seleccionados son:

Choices: Representa una lista de alternativas posibles que el usuario pronunciará dentro de las restricciones de una gramática de reconocimiento de voz.
GrammarBuilder: Proporciona un mecanismo para construir las restricciones de una gramática de reconocimiento de voz, permitiendo armar una gramática a partir de un conjunto de frases y opciones

(Choices). De esta manera se puede definir la forma en que las palabras pueden ser combinadas para ser entendidas por el reconocedor. Con estas restricciones al motor de voz, se logra que se realicen mejores escogencias entre sonidos ambiguos, incrementando así la precisión del sistema.
Grammar: Proporciona soporte en tiempo de ejecución para la obtención y gestión de la información de una gramática de reconocimiento de voz.

De esta manera, se predefinió en el aplicativo por ejemplo, el reconocer las órdenes de domótica que se muestran en la Figura 14. En donde se dan como resultados válidos frases como “Prender Luz Pasillo”, “Prender Luz Entrada”, “Apagar Luz Pasillo”, “Apagar Luz Alcoba”, “Abrir Puerta Alcoba”, “Cerrar Cortina Entrada”, “Abrir Cortina Alcoba”, entre otras.

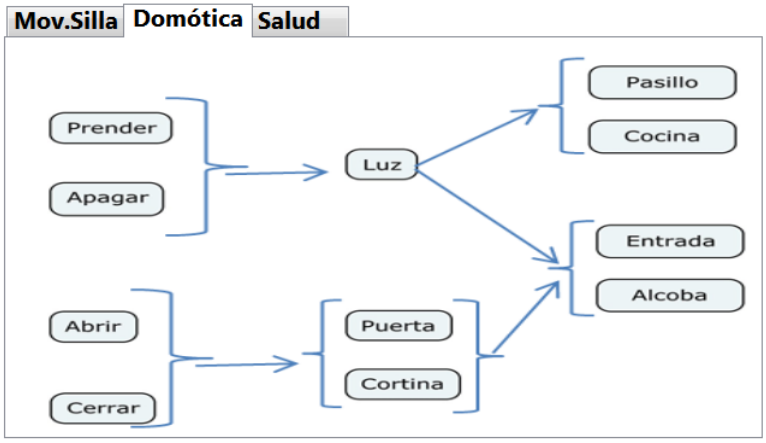

Figura 14: Secuencia de posibles palabras para formar frases relacionadas con órdenes de domótica.

A continuación, se muestra la porción de código con las clases del System.Speech.Recognition para especificar la gramática correspondiente a las órdenes de domótica que se acaban de mencionar.

```
//Gramática para comandos de Domótica:
            GrammarBuilder ON_OFF = new GrammarBuilder(new Choices("prender", "apagar"));
            GrammarBuilder EST2 = new GrammarBuilder(new Choices("Abrir", "Cerrar"));
            GrammarBuilder UBICA1 = new GrammarBuilder(new Choices("Pasillo", "Cocina"));
            GrammarBuilder UBICA2 = new GrammarBuilder(new Choices("Alcoba", "Entrada"));
            GrammarBuilder UBICACIONES = new GrammarBuilder(new Choices(UBICA1, UBICA2));
            GrammarBuilder ELEM = new GrammarBuilder(new Choices("Puerta", "Cortina"));

            GrammarBuilder DOMOTIC1 = new GrammarBuilder();
            DOMOTIC1.Append(ON_OFF);
            DOMOTIC1.Append("Luz");
            DOMOTIC1.Append(UBICACIONES);

            GrammarBuilder DOMOTIC2 = new GrammarBuilder();
            DOMOTIC2.Append(EST2);
            DOMOTIC2.Append(ELEM);
```

```
            DOMOTIC2.Append(UBICA2);

// Creación de una instancia del Grammar con los objetos del GrammarBuilder.
//y cargarla dentro del speech recognizer.
Grammar g2 = new Grammar(new Choices(DOMOTIC1, DOMOTIC2));
            g2.Name = "Domotica";
            sr.LoadGrammar(g2);
```

Las diferentes opciones que pueden ir en un mismo nivel de la frase se agrupan mediante una clase **Choices**, así por ejemplo, las opciones “prender” o “apagar”, están contenidas en la misma instancia del *Choices* que hace parte del **GrammarBuilder** denominado *ON_OFF*. Ya para establecer la secuencia en la que las palabras definidas en los *Choices* aparecerán para formar frases, se definieron dos *GrammarBuilder* llamados *DOMOTIC1* y *DOMOTIC2*, que como se observa por ejemplo en *DOMOTIC1*, por medio del método **Append(GrammarBuilder)** se genera la secuencia donde con su respectivo orden primero va el *GrammarBuilder ON_OFF* (que contiene "prender" y "apagar"), le sigue la palabra “luz” y finalmente se le adiciona el *GrammarBuilder UBICACIONES* que contiene a su vez los *GrammarBuilder UBICA1* (con las opciones "Pasillo" y "Cocina"), y *UBICA2* (con "Alcoba" y "Entrada"). Aceptando de esta manera por ejemplo la frase “Prender luz entrada”. Similar sucede con la contrucción de *DOMOTIC2* que acepta por ejemplo la frase “Cerrar cortina alcoba”, pero no acepta “Cerrar Luz Cocina” ya que no se definió en su construcción. Finalmente, toda las posibles frases que se pueden obtener para el vocabulario referente a órdenes de domótica, se agrupan en una sola gramatica mediante la clase **Grammar** que se definió como g2 y que se compone por las opciones presentadas por el *GrammarBuilder DOMOTIC1* y el *GrammarBuilder DOMOTIC2*.

Para cargar en la aplicación cada una de las gramáticas definidas, siendo el caso de las relacionadas con órdenes de domótica la que se llamo como g2, se utiliza el método **LoadGrammar()** quien realiza la carga sincrónicamente; éste método pertenece a la clase *SpeechRecognitionEngine*, quien a su vez pertenece al espacio de nombre System.Speech.Recognition y proporciona acceso para ejecutar cualquier servicio de reconocimiento de voz correctamente instalado en un sistema de escritorio de Windows.

Con un procedimiento similar al anterior, se definió cada una de las gramáticas correspondientes para los diferentes comandos que contiene la aplicación desarrollada. Las gramáticas prefijadas corresponden a los comandos de movimiento de la silla, comandos relacionados con domótica,

comandos para la toma de signos vitales, comandos para el desplazamiento por las pestañas de la aplicación y para la activación de sus principales botones.

La gramática que se actualiza en tiempo de ejecución y que es configurable por el usuario se refiere a la perteneciente para asociar un comando con un destinatario de correo electrónico y para controlar la apertura de las aplicaciones predefinidas instaladas en el computador del usuario.

Las posibles frases que la aplicación está en posibilidad de reconocer, se listan a continuación:

- Para comandos relacionados con la toma de signos vitales, el usuario puede pronunciar como primera palabra las opciones “Capturar”, “Medir” o “Tomar”; y como segunda palabra las opciones de “Electro”, “Presión”, “Temperatura”, “pulso”, u “oxígeno”. Formando de esta manera frases válidas como “Tomar presión” o “Capturar electro”. Los comandos de salud se definieron según los módulos de transmisión de signos vitales que Docentes del proyecto de silla de ruedas inteligente están elaborando y que serán acoplados para responder en un proyecto futuro a dichos comandos.

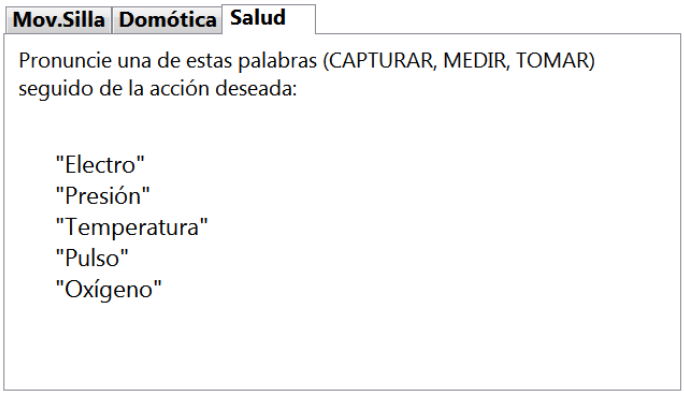

Figura 15: visualización de los comandos que el usuario puede pronunciar referente a la toma de signos vitales.

- Para comandos relacionados con el movimiento de la silla, el usuario debe pronunciar como primera palabra “Mover”, esto pensando en una palabra de seguridad que anteceda a la orden del movimiento como tal; como segunda palabra podrá pronunciar alguna de las acciones deseadas “Izquierda”, “Adelante”, “Derecha”, “Atrás”, “Parar”, “lento”, o “suave”. Formando de esta manera frases válidas como “Mover adelante” o “mover lento”. Los comandos del movimiento de la silla se definieron con las mismas opciones que posee el Joystick de la silla de ruedas eléctrica que adquirió la UAM para el grupo de investigación que trabaja en el

proyecto de silla de ruedas inteligente, al igual que se colocaron las imágenes de la tortuga para “lento” y de liebre para “rápido” buscando una semejanza con las gráficas que trae dicho Joystick.

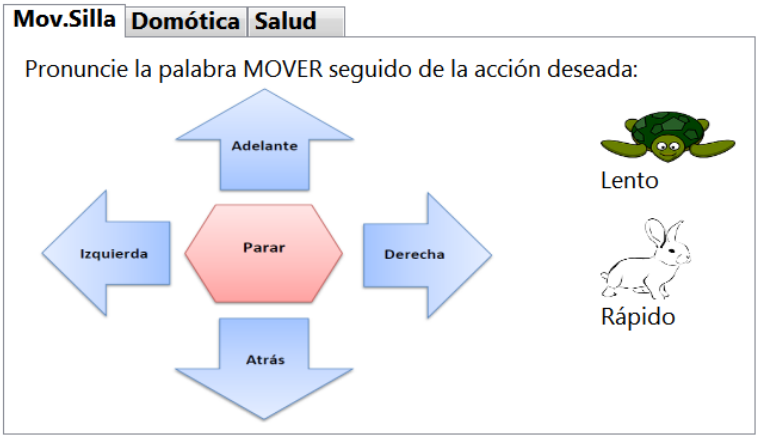

Figura 16: Visualización de los comandos que el usuario puede pronunciar referente al movimiento de la silla.

- Para comandos relacionados con el desplazamiento por las pestañas de la aplicación, el usuario debe pronunciar como primera palabra “Ver” y como segunda palabra las opciones de “Silla”, “Domótica”, “Salud”, “Inicio”, “Configuración”, “programas”, o “Pruebas”. Formando de esta manera frases válidas como “Ver domótica” o “Ver Programas”. Se debe aclarar que la pestaña denominada Pruebas solo se tiene para visualizar las variables del estado de reconocimiento, pero para una versión para el usuario final esta pestaña no se incluiría. Adicionalmente no se tiene el desplazamiento por voz para la pestaña “Correo a enviar” ya que esta solo actualiza sus campos cuando el usuario pronuncia el comando asociado para enviar un correo electrónico a algún destinatario (función que se explica más abajo), momento en el que automáticamente se despliega dicha pestaña y no tiene sentido visualizarla cuando no hay un destinatario asociado.

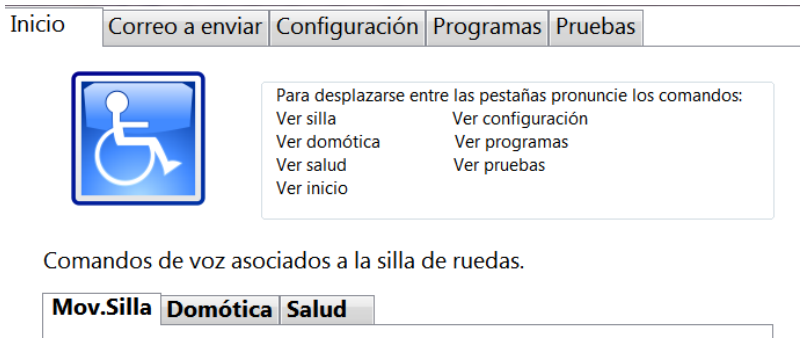

Figura 17: Visualización de los comandos que el usuario puede pronunciar referente al desplazamiento por las pestañas de la aplicación.

- Los comandos relacionados con órdenes de domótica, ya se mencionaron más arriba cuando se explicó la porción de código para generar una gramática.

- Para comandos relacionados con la activación de los principales botones de la aplicación, el usuario debe pronunciar la frase “voz detener” o “Voz activar” para detener el reconocimiento o volverlo a activar respectivamente. El botón para detener el reconocimiento es muy importante en la aplicación, una vez que si el usuario desea establecer una conversación con alguien o pronunciar palabras que no van destinadas a los comandos por voz, entonces deshabilita el sistema de reconocimiento para que éste no acepte comandos en dicho momento. Una vez el usuario desea restablecer el reconocimiento, activa el botón llamado “Activar de nuevo reconocimiento”. El otro botón controlado por voz es el que permite enviar correo para alguno de los destinatarios que previamente ha configurado el usuario (este se encuentra dentro de la pestaña “correo a enviar”) y para activarlo se debe pronunciar la frase “Enviar Correo”.

Figura 18: Visualización de los botones que el usuario puede activar por comandos de voz.

- Una de las dos gramáticas configurables por el usuario en la aplicación y que se actualiza en tiempo de ejecución es la referente a asociar un comando con un destinatario de correo electrónico. Para el ingreso de los datos se han habilitado dos campos, unos para digitar el comando con el que asociará una dirección de correo electrónico de destinatario y el otro para digitar como tal la dirección del correo. Una vez la aplicación tiene configurados los correos

que el usuario ha predefinido, solo se debe pronunciar como primera palabra “Correo” y como segunda palabra la referente a uno de los comandos que se encuentre en la lista de correos ingresado por el usuario. De esta manera se conformarán frases válidas como “Correo hermano” o “Correo vecino”, conservándose siempre la palabra “correo” y variando la segunda palabra según lo defina el usuario.

El proceso de generación de la gramática es similar al código descrito anteriormente para órdenes de domótica, teniendo en cuenta que se agrega una nueva instancia a la clase **Choices** de dicha gramática por cada comando que el usuario almacene.

A la aplicación se le involucró la función de enviar por comandos de voz correos electrónicos predefinidos por el usuario, aprovechando que el programa corre en un computador y pensando en la posibilidad de que cuando se esté en una emergencia o por algún motivo requiera comunicarse con alguien que ya este configurado como destinatario, solo sea pronunciar el comando correspondiente para que le llegue el correo con la plantilla predefinida a dicha persona.

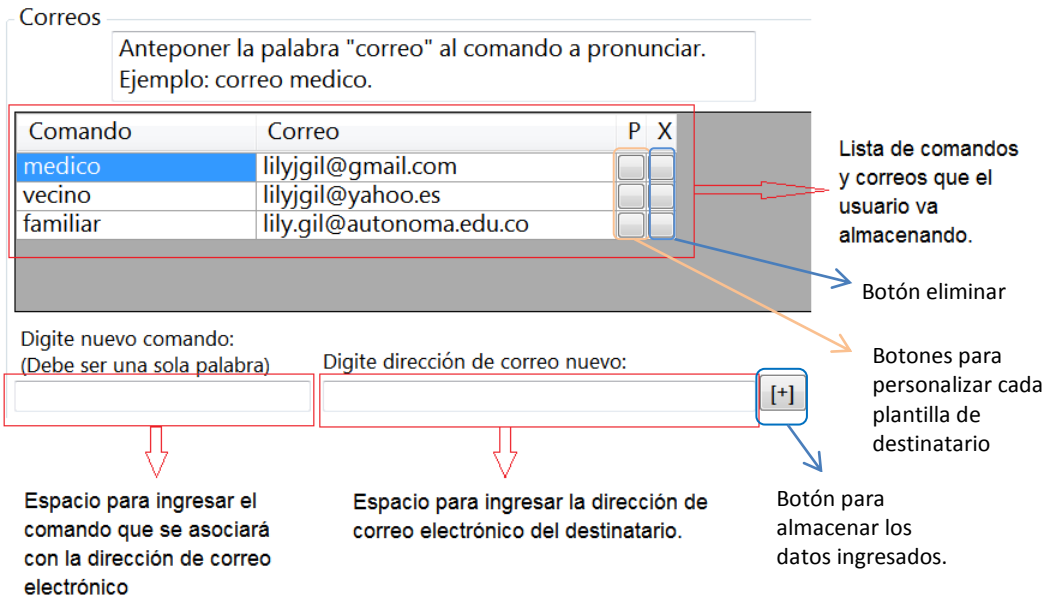

Figura 19: Visualización del entorno en el que el usuario configura los comandos asociados a cuentas de correo electrónico de destinatarios deseados.

- La otra gramática configurable por el usuario en la aplicación y que se actualiza en tiempo de ejecución es la referente a la apertura de aplicaciones instaladas en el computador en el que

corre la aplicación. El proceso es similar al efectuado para el envío de correos electrónicos a destinatarios específicos, descrito anteriormente. Para el ingreso de los datos también se han habilitado dos campos, uno para digitar el comando que se asociará a la ruta del ejecutable de una aplicación específica y el otro para ingresar como tal la ruta de ese ejecutable. Una vez se tienen almacenados los comandos que lanzarán las aplicaciones, solo se debe pronunciar como primera palabra “aplicación” y como segunda palabra la referente a uno de los comandos que se encuentre en la lista de aplicaciones que el usuario predefinió. De esta manera se conformarán frases válidas como “aplicación calculadora” o “aplicación fecha”, conservándose siempre la palabra “aplicación” y variando la segunda palabra según lo defina el usuario. Esta función se incluyó, pensando a futuro en que se puede acoplar con programas que tan pronto se lancen empiecen a capturar datos del usuario de la silla.

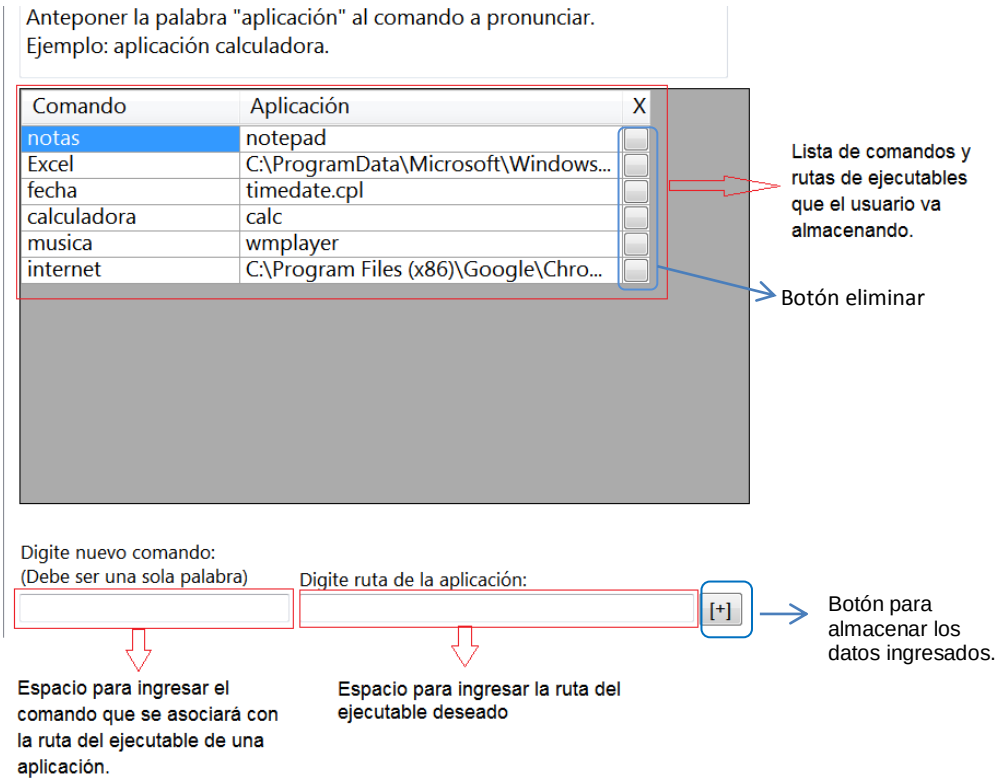

Figura 20: Visualización del entorno en el que el usuario configura el comando asociado a una ruta del ejecutable de una aplicación deseada.

3.5 Interfaz de usuario

3.5.1. Aspecto general, retroalimentación de la respuesta del reconocedor e indicaciones para los comandos a pronunciar

La aplicación se desarrolló bajo una interfaz compuesta principalmente por pestañas a las que se puede acceder por comandos de voz. Está dividida en una sección fija (lado derecho) y una sección que se intercambia (lado izquierdo) según la pestaña que se esté visualizando en el momento.

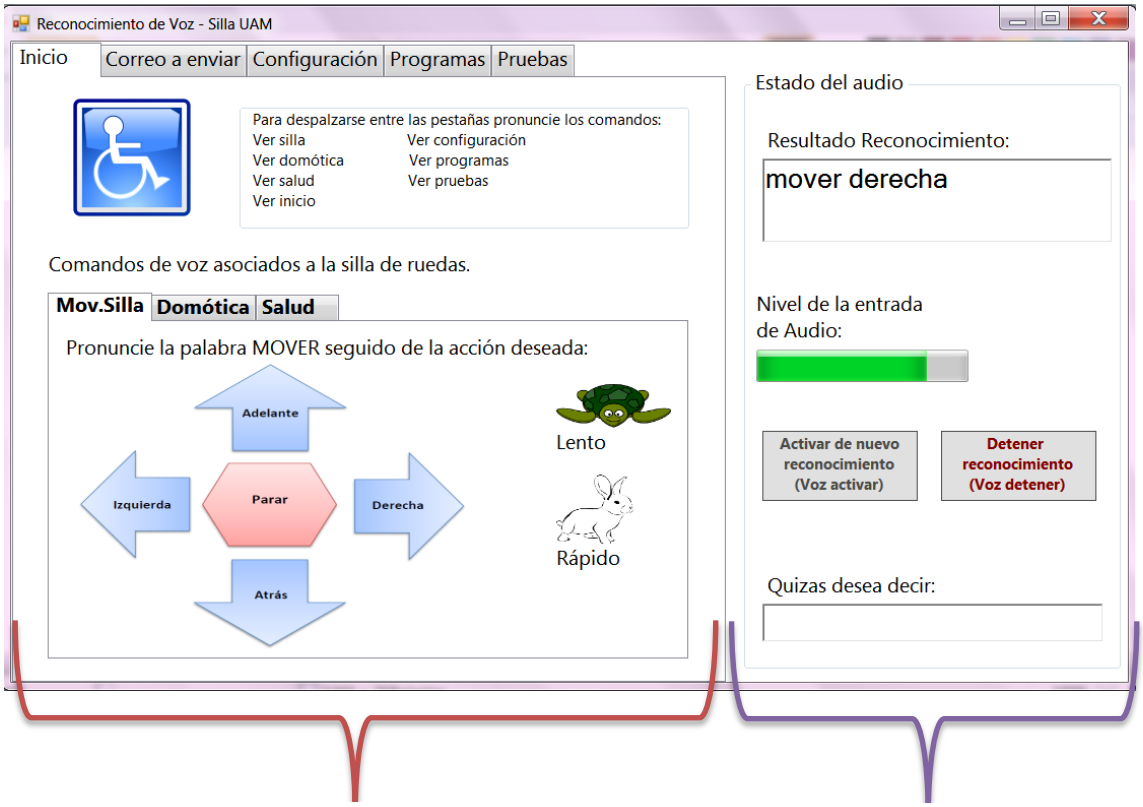

Figura 21: Distribución de la interfaz con una sección intercambiable y una sección fija.

La sección fija contiene información de interés para el usuario en cualquiera de las pestañas dentro de las que se encuentre y por eso siempre permanece visible en la aplicación; Dicha información se refiere al resultado del reconocimiento (retroalimentación visual del comportamiento del reconocedor), una barra indicadora de entrada de audio, los botones para

desactivar y activar de nuevo el reconocimiento y un cajón que indica en caso de no ser exitoso el reconocimiento, las frases que más se acercan a lo que el reconocedor alcance a capturar. Cuando una frase es reconocida con éxito, la aplicación reproduce como medio de retroalimentación auditiva la palabra “Éxito”.

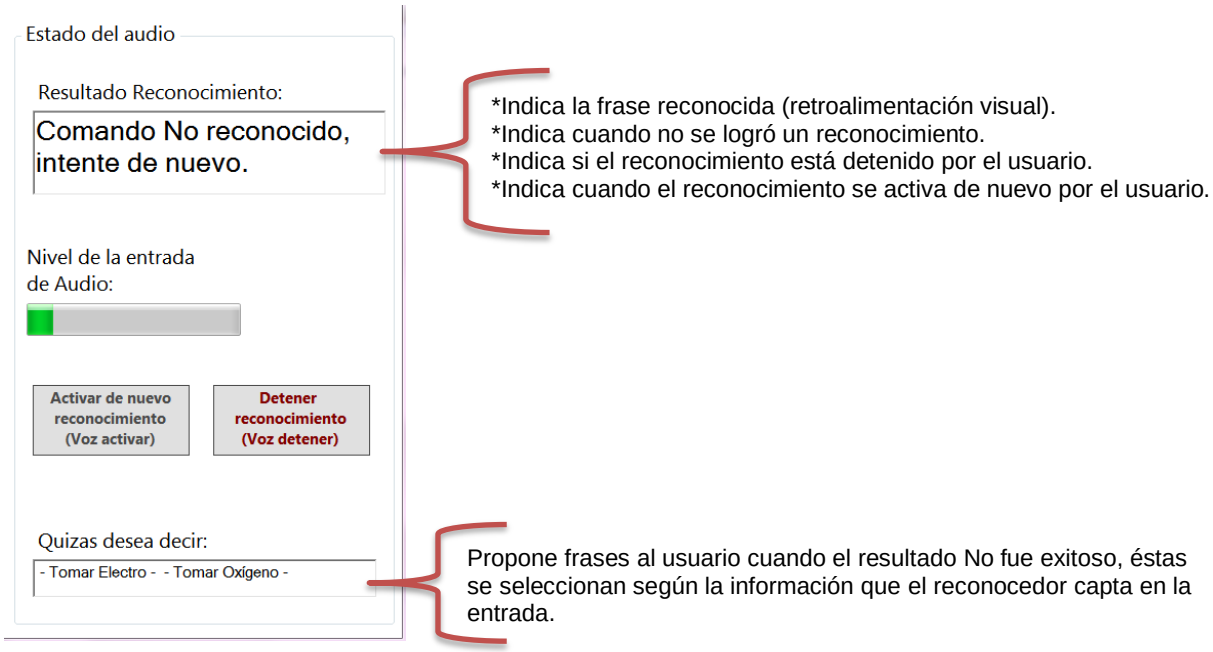

Figura 22: sección fija de la interfaz de usuario.

La pestaña relacionada con comandos de movimiento de la silla, se construyó con elementos iconográficos que se asemejaran al joystick original con el que viene la silla adquirida por la UAM, tal como se puede observar en la sección punteada de la Figura 23.

Para recordar los comandos a pronunciar referentes al movimiento de la silla, a órdenes de domótica y a toma de signos vitales cada una de las respectivas pestañas muestra al usuario las palabras que debe pronunciar en su orden para formar las respectivas frases, como se observa en el recuadro punteado rojo de la Figura 23, Figura 24 y Figura 25. Así mismo, los botones de activar y desactivar reconocimiento indican el comando de voz que los controla y en la pestaña Inicio se encuentra la información de los comandos para desplazarse por las diferentes pestañas, ver recuadro puenteado rojo de la Figura 26. Las pestañas denominadas “Correo a enviar”, “Configuración” y “Programas” también indican la forma correcta en que se deben decir las frases

correspondientes a los comandos que se quieren reconocer, ver recuadro puenteado rojo de la figura Figura 27, Figura 28 y Figura 29.

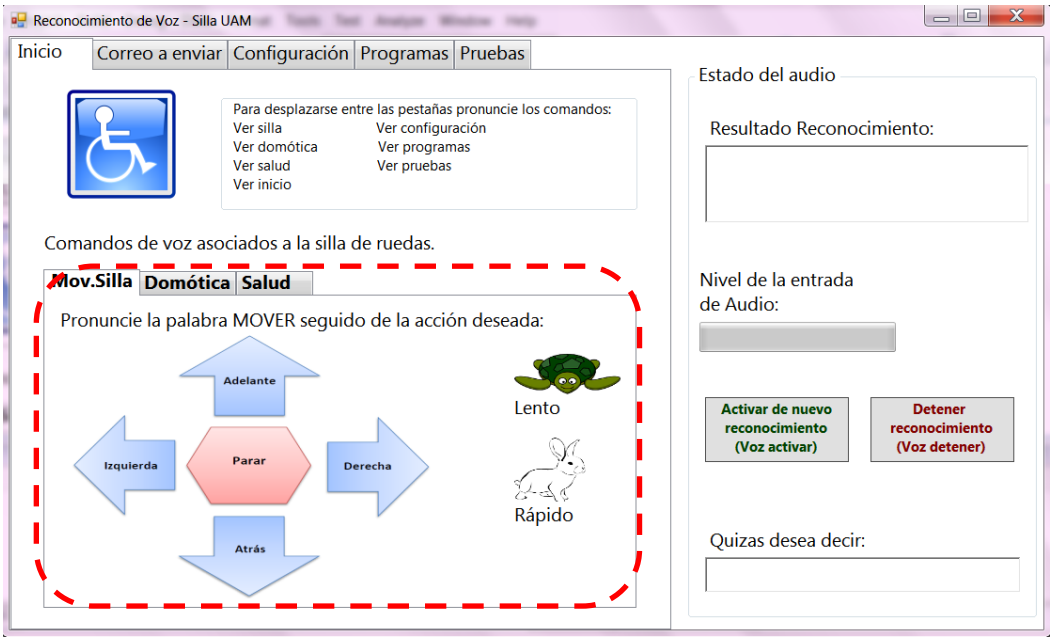

Figura 23: Comandos a pronunciar relacionados con el movimiento de la silla

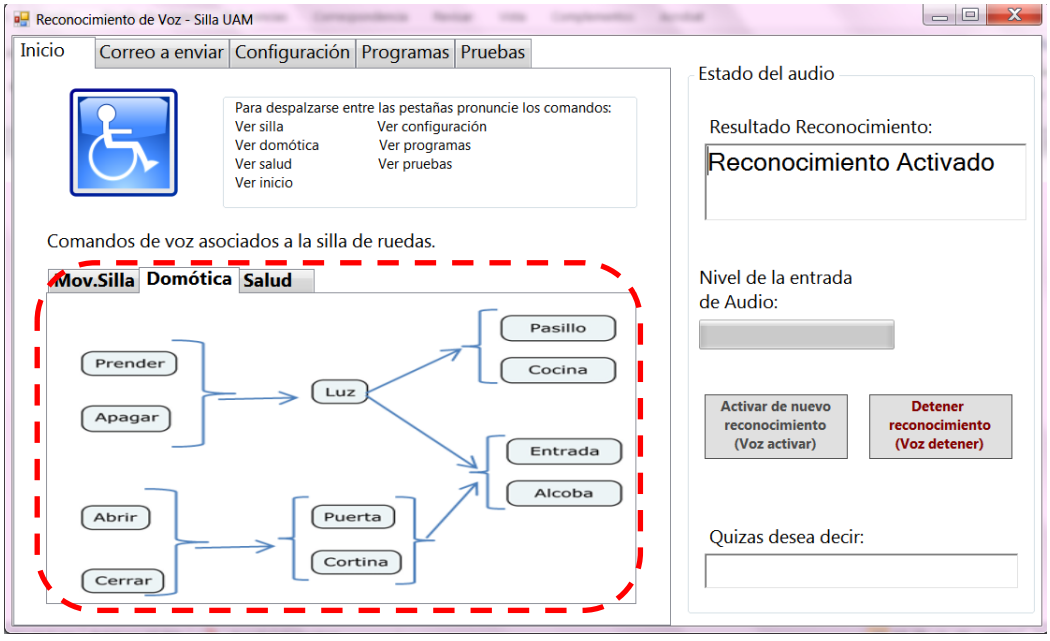

Figura 24: Comandos a pronunciar relacionados con órdenes de domótica

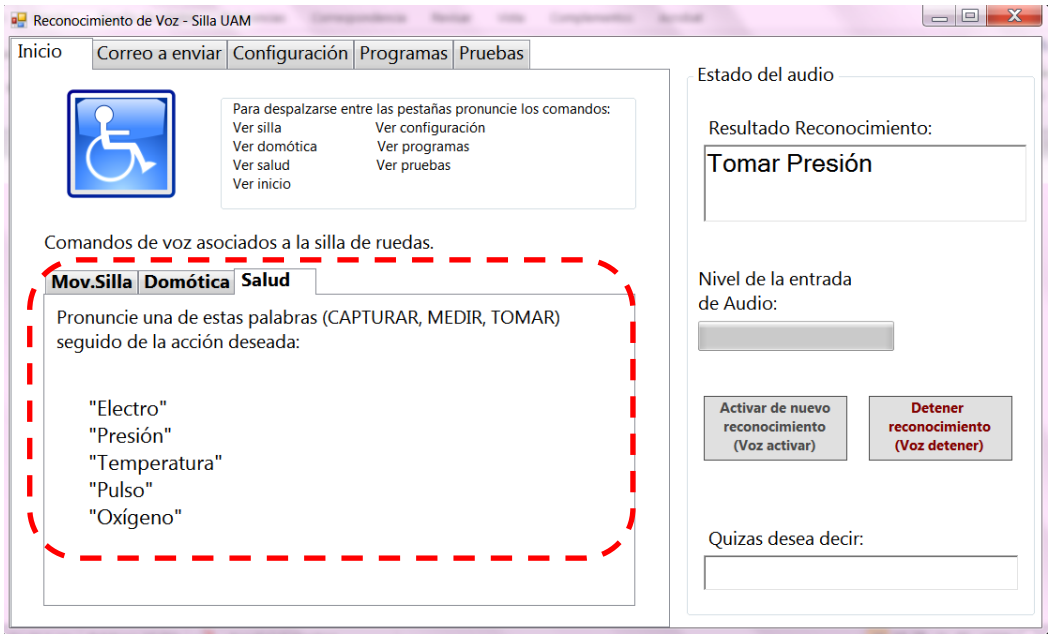

Figura 25: Comandos a pronunciar relacionados con la toma de signos vitales

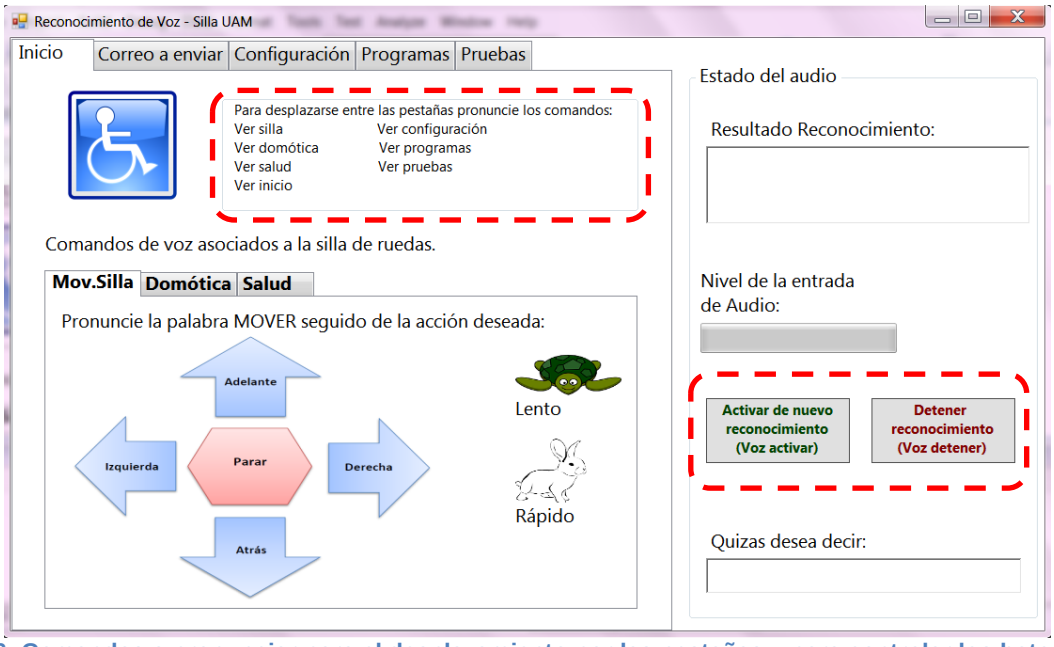

Figura 26: Comandos a pronunciar para el desplazamiento por las pestañas y para controlar los botones de la activación del reconocimiento.

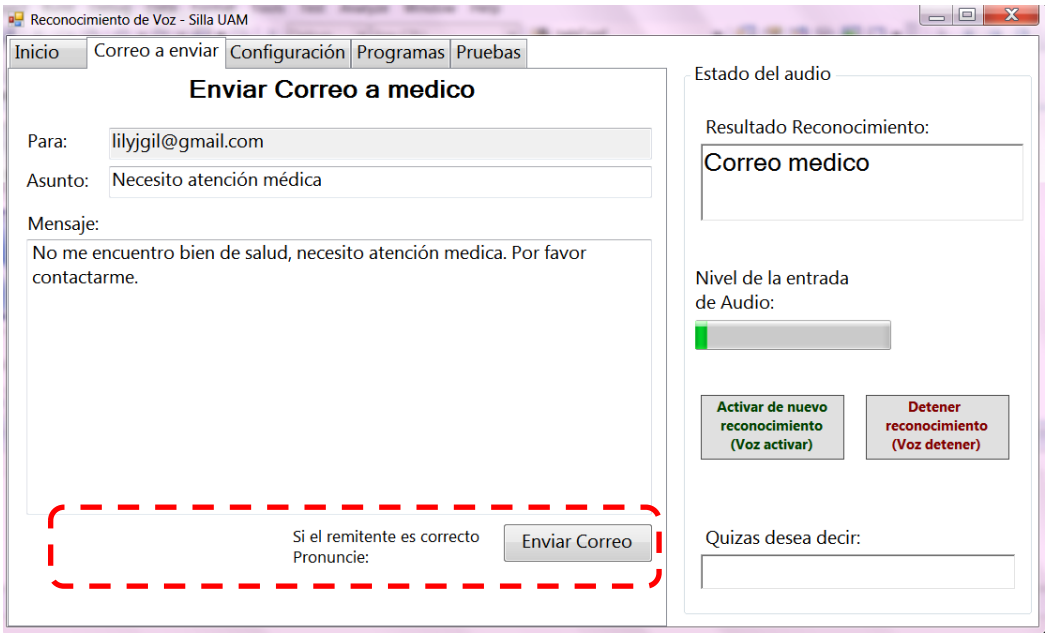

Figura 27: Comando a pronunciar para enviar correo.

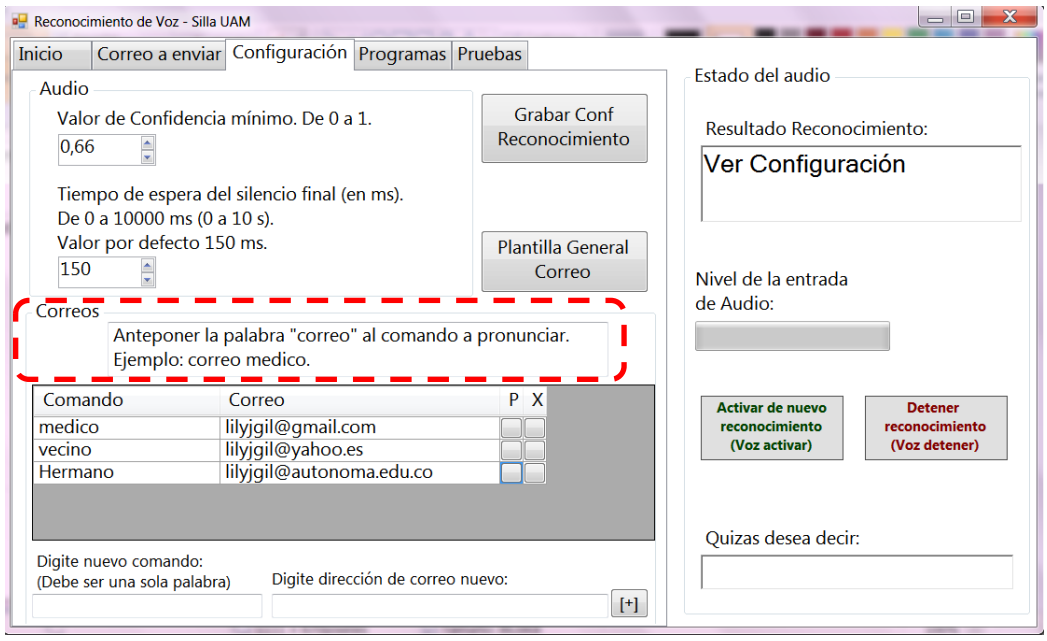

Figura 28: Comandos a pronunciar para seleccionar destinatario de correo.

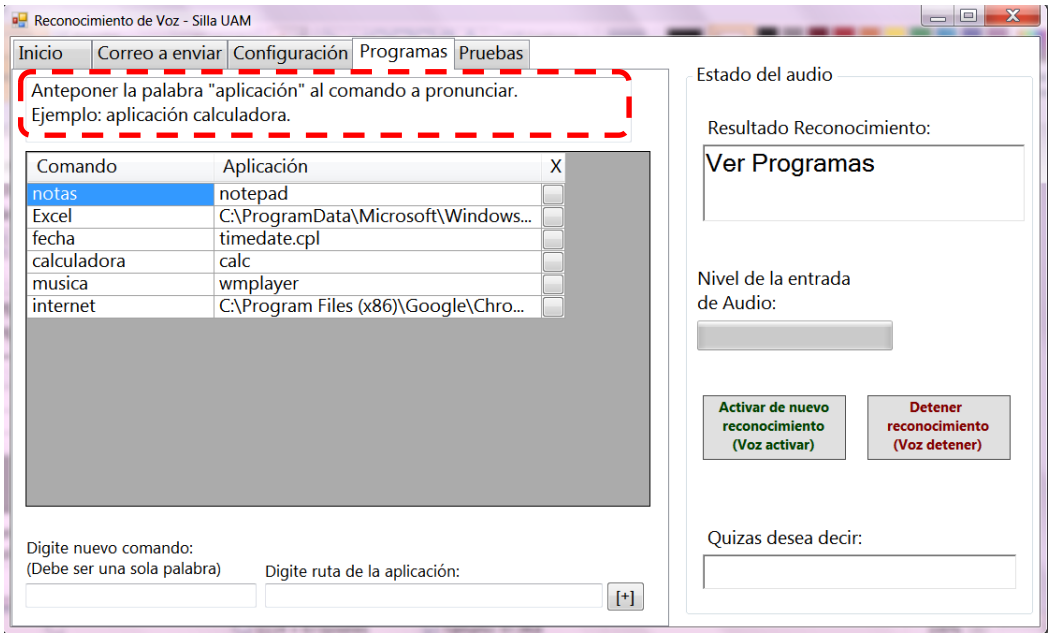

Figura 29: Comandos a pronunciar para lanzar aplicaciones.

3.5.2. Configuraciones de la aplicación que no se realizan por comandos de voz:

Hay algunas configuraciones que el usuario o asistente del usuario (en caso de que este no pueda usar sus manos) debe realizar para que el paciente disfrute de la aplicación, esas configuraciones son:

- Establecer el valor de confidencia mínimo que mejor responde al entorno y a la voz del usuario, esta propiedad le da una restricción de nivel de confianza al reconocedor, si el valor es muy bajo, puede detectar erróneamente una mayor cantidad de palabras como válidas, y si es muy alto puede bloquear una mayor cantidad de frases que si son correctas y tomarlas como no válidas; también se debe establecer el tiempo de espera del silencio final, este básicamente fija el tiempo que la aplicación debe esperar para detectar las frases una vez el usuario las ha pronunciado. Con un valor muy alto se pierde la respuesta en tiempo real del sistema. Cada que se modifique el valor de confidencia mínimo o el tiempo de espera final, se debe dar clic en el botón denominado “Grabar Configuración Reconocimiento”. La Figura 30 muestra en los espacios con borde rojos punteado dichos elementos en la aplicación.

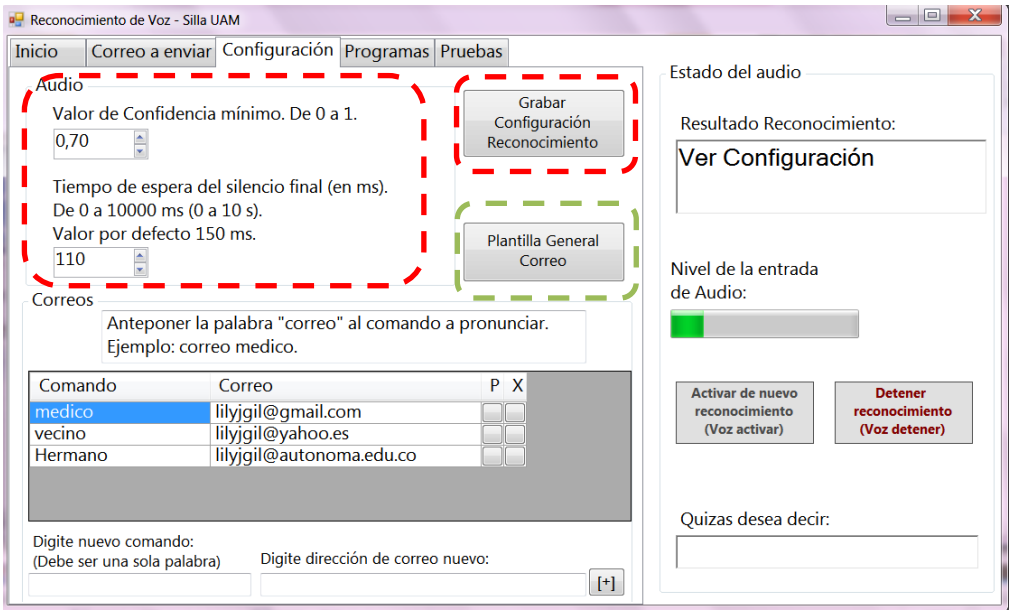

Figura 30: Configuración de parámetros del reconocedor.

- Al dar clic en el botón denominado “Plantilla General Correo”, borde verde punteado en la Figura 30, se accede a una nueva ventana en donde se debe configurar la dirección de correo electrónico del remitente de los mensajes (usuario de la aplicación), se debe proporcionar la clave de acceso a dicho correo para poder enlazarlo con la aplicación, configurar el nombre que aparecerá como remitente en los correos y configurar un asunto y mensaje por defecto que se cargará como plantilla inicial en todos los correos de destinatario que se ingresen hasta que éstos sean personalizados ingresando a la plantilla particular de cada uno de ellos. La Figura 31 muestra dicha plantilla.

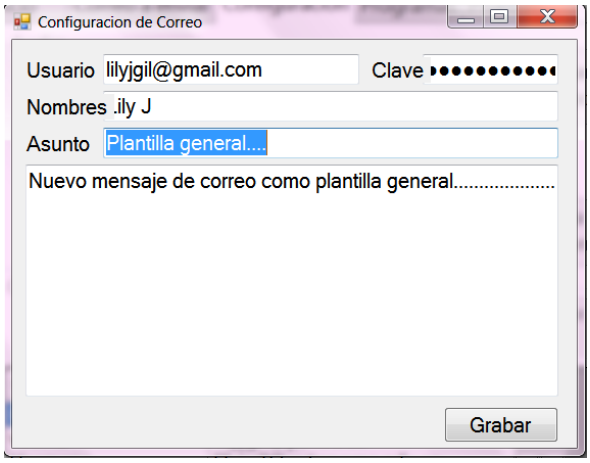

Figura 31: Plantilla para configurar datos del remitente

- Cada que se ingresa un comando y su correo asociado para un destinatario deseado, se debe dar clic en el botón que almacenará dichos datos y los enviará a la lista visible de correos ya configurados. Los botones para eliminar algún dato de la lista de correos o para personalizar las plantillas, también deben ser accedidos dando clic sobre ellos. La Figura 19 muestra estas opciones.
- Para personalizar cada plantilla de destinatario se debe dar clic en el botón indicado en la Figura 19 para tal fin, la plantilla que se despliega se muestra en la Figura 32, los campos a modificar allí son el asunto y el mensaje, este último no debe ser ingresado con saltos de línea (tecla Enter).

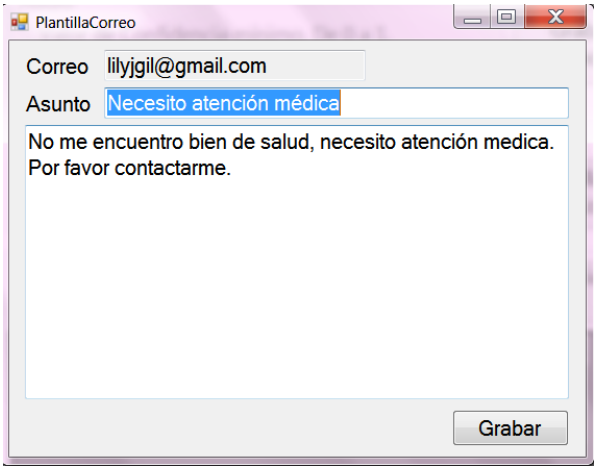

Figura 32: Plantilla para personalizar el mensaje de cada destinatario.

- Cada que se ingresa un comando y la ruta asociada que abre una aplicación deseada, se debe dar clic en el botón que almacenará dichos datos y los enviará a la lista visible de aplicaciones ya configuradas. Los botones para eliminar algún dato de la lista de aplicaciones también deben ser accedidos dando clic sobre ellos. La Figura 20 muestra estas opciones.

La pestaña denominada “Pruebas”, como ya se mencionó, se eliminaría del aplicativo para un usuario final. Esta pestaña tiene validez durante el de desarrollo de la aplicación ya que permite monitorear el comportamiento del reconocedor de voz.

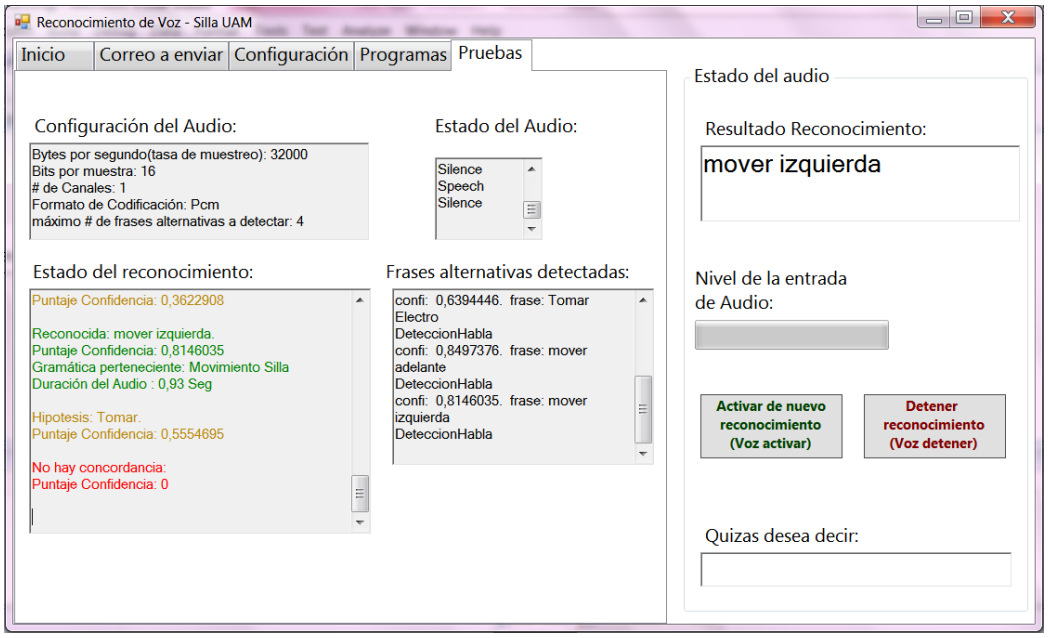

Figura 33: Monitoreo de la respuesta del reconocedor.

4. PRUEBAS, RESULTADOS Y DISCUSION

4.1 Pruebas para determinar el nivel de desempeño obtenido en la aplicación

Para determinar el nivel de desempeño obtenido en la aplicación para el reconocimiento de los comando de voz en español para un vocabulario cerrado e independiente del hablante, se realizaron pruebas que involucraron niveles de ruido en el ambiente en tres rangos diferentes acorde con la legislación Colombiana sobre estándares máximos permisibles de niveles de ruido ambiental, así como participaron en la prueba igual número de hombres que de mujeres.

- Condiciones del ruido ambiental:

Las pruebas para hombres y mujeres se realizaron bajo los mismos 3 rangos de niveles de ruido medidos en dB(A). El decibelio (dB) es una medida de nivel de presión sonora que se utiliza en acústica para cuantificar la “cantidad” de ruido existente. A mayor cantidad, mayor molestia. Es una magnitud de tipo logarítmica. Así mismo el Decibelio con ponderación A dB(A) es una unidad de nivel sonoro medido con un filtro previo que quita parte de las bajas y las muy altas frecuencias adaptándose a la percepción del oído humano. De esta manera, después de la medición se filtra el sonido para conservar solamente las frecuencias más dañinas para el oído, razón por la cual la exposición medida en dB(A) es un buen indicador del riesgo auditivo y es la unidad más utilizada para la medición de los niveles de ruido ambiental. [62]. Según la Organización Mundial de la Salud, “El ruido urbano (también denominado ruido ambiental, ruido residencial o ruido doméstico) se define como el ruido emitido por todas las fuentes a excepción de las áreas industriales. Las fuentes principales de ruido urbano son tránsito automotor, ferroviario y aéreo, la construcción y obras públicas y el vecindario. Las principales fuentes de ruido en interiores son los sistemas de ventilación, máquinas de oficina, artefactos domésticos y vecinos.” [63]

Las pruebas realizadas se establecieron con rangos de nivel de ruido ambiental según la vigente resolución 0627 del 7 de Abril de 2006 del entonces Ministerio de ambiente, vivienda y desarrollo territorial por la cual se establece la norma nacional de emisión de ruido y ruido ambiental, el cual en el capítulo III denominado “del ruido ambiental” indica los estándares máximos permisibles de niveles de ruido ambiental, expresados en decibeles dB(A) y el cual se muestra en la Tabla 2.

Sector	Subsector	Estándares máximos permisibles de niveles de ruido ambiental en dB(A)	
		Día	Noche
Sector A. Tranquilidad y Silencio	Hospitales, bibliotecas, guarderias, sanatorios, hogares geriátricos.	55	45
Sector B. Tranquilidad y Ruido Moderado	Zonas residenciales o exclusivamente destinadas para desarrollo habitacional, hotelería y hospedajes.	65	50
	Universidades, colegios, escuelas, centros de estudio e investigación		
	Parques en zonas urbanas diferentes a los parques mecánicos al aire libre		
Sector C. Ruído Intermedio Restringido	Zonas con usos permitidos industriales, como industrias en general, zonas portuarias, parques industriales, zonas francas.	75	70
	Zonas con usos permitidos comerciales, como centros comerciales, almacenes, locales o instalaciones de tipo comercial, talleres de mecánica automotriz e industrial, centros deportivos y recreativos, gimnasios, restaurantes, bares, tabernas, discotecas, bingos, casinos.	70	55
	Zonas con usos permitidos de oficinas.	65	50
	Zonas con usos institucionales.		
	Zonas con otros usos relacionados, como parques mecánicos al aire libre, áreas destinadas a espectáculos públicos al aire libre, vías troncales, autopistas, vías arterias, vías principales.	80	70
Sector D. Zona Suburbana o Rural de Tranquilidad y Ruído Moderado	Residencial suburbana.	55	45
	Rural habitada destinada a explotación agropecuaria.		
	Zonas de Recreación y descanso, como parques naturales y reservas naturales.		

Tabla 2: Estándares máximos permisibles de niveles de ruido ambiental, expresados en decibeles dB(A)

Para asegurar que las pruebas con todos los usuarios se realizaran dentro del mismo rango de dB(A), se utilizó como instrumento de medida un sonómetro marca UNI-T, referencia UT352. El cual tiene un rango de medición entre 30dB y 130dB, con una exactitud de ±1.5dB. El sonómetro se configuró para medir con el filtro de ponderación frecuencial A y el filtro de ponderación temporal F (Rápido) que tiene un tiempo de respuesta de 125 ms. El medidor utilizado se muestra en la Figura 34 .Las mediciones se realizaron por cada uno de los usuarios y para cada uno de los tres niveles de ruido deseado. Esta medición se efectuó justo junto al micrófono que el usuario por medio de una diadema ya tenía ubicado cerca de su boca, tomando así el valor de dB(A) que se estaba percibiendo alrededor del micrófono.

Figura 34: Sonómetro utilizado en las pruebas

La prueba con el primer rango de ruido ambiental se realizó en un espacio cerrado silencioso alejado del tráfico vehicular, con mediciones en el sonómetro comprendidas entre un mínimo de 35 dB(A) y un máximo de 55 dB(A). Para esta prueba no se adicionó ningún ruido externo, los valores eran los que se percibían en el lugar del micrófono tratando de que se conservara el silencio. Estas mediciones se encuentran dentro del rango de un sector A. Tranquilidad y Silencio según Tabla 2.

Para las pruebas dos y tres se conservó el mismo sitio donde se hizo la prueba uno pero se le adicionó ruido al lugar. Para generar este ruido se utilizó un computador en el que se corrió la aplicación online myNoise™.net. Esta aplicación ofrece una colección de generadores de ruido online que cubre todo el rango de frecuencia audible desde los 20 Hz hasta los 20 KHz y cada ruido seleccionado se puede calibrar según necesidades del usuario. Para las pruebas realizadas se seleccionó el generador de ruido de fondo llamado Coffee-Shop, que simula el ruido que se siente en una cafetería concurrida donde hay sonido de cubiertos, de objetos retumbando, de personas charlando, murmurando y tosiendo, entre otros sonidos. El ruido seleccionado se puede escuchar ingresando a [64]. Como los parlantes del computador no alcanzaban a generar los niveles de decibeles requeridos para la prueba, se conectó un parlante externo al mismo donde controlando el nivel de volumen del parlante se podía controlar los rangos de dB(A) deseados para la prueba.

La prueba dos se estableció en un rango de dB(A) desde un valor mínimo de 60dB(A) hasta un valor máximo de 72 dB(A) y la prueba tres se estableció en un rango desde un valor mínimo de 73dB(A) hasta un valor máximo de 85 dB(A), dichos rangos se mantuvieron controlados con el volumen adecuado del parlante y con las mediciones del sonómetro. Según la Tabla 2, para un sector de ruido Intermedio Restringido, el valor máximo permisible de nivel de ruido ambiental que se encuentra en la legislación se establece en 80dB(A) y pertenece al caso de Zonas como parques mecánicos al aire libre, áreas destinadas a espectáculos públicos al aire libre, vías troncales, autopistas, vías arterias, vías principales. Partiendo de dicha tabla, los rangos para las pruebas dos y tres se encuentran en el Sector C. Ruido Intermedio Restringido; y específicamente la prueba tres incluye mediciones superiores a los 80dB(A) que sería el ambiente más ruidoso con el que el usuario se encontraría como ruido ambiental según la resolución 0627 del 7 de Abril de 2006. La Tabla 3 muestra los rangos definidos para los tres tipos de pruebas realizadas a igual número de hombres y de mujeres.

	Rango dB(A)	**Ruido**
Prueba # 1	35 dB(A) hasta 55 dB(A)	Lugar en silencio
Prueba # 2	60 dB(A) hasta 72 dB(A)	Adicionando ruido tipo Coffee-Shop.
Prueba # 3	73 dB(A) hasta 85 dB(A)	

Tabla 3: Rangos de ruido establecidos para las tres pruebas.

- Entorno de la prueba:

Las pruebas fueron realizadas con 10 hombres y 10 mujeres, cada persona debía realizar la prueba en los tres ambientes y debía colocarse un micrófono diadema a una distancia de aproximadamente 3 cm de la boca al momento de pronunciar los comandos. Los participantes todos Colombianos poseen edades entre los 16 y los 70 años.

En las pruebas con los rangos dos y tres donde se debía generar ruido con la ayuda de un parlante externo, se aseguró que la distancia para todos los participantes fuera de 1m entre la fuente generadora del ruido (parlante) y el micrófono posicionado cerca a la boca del interlocutor, así como se graduaba el volumen del parlante hasta que el sonómetro marcara los rangos establecidos para cada prueba.

El Computador en el que se realizaron las pruebas corre bajo un sistema operativo Windows 7 con Service Pack 1 de 64 bits, memoria RAM de 4GB y procesador Intel Core i5. La diadema micrófono

usada es marca Genius HS-210U con conexión por USB. La Figura 35 muestra las condiciones en que se realizaron las pruebas.

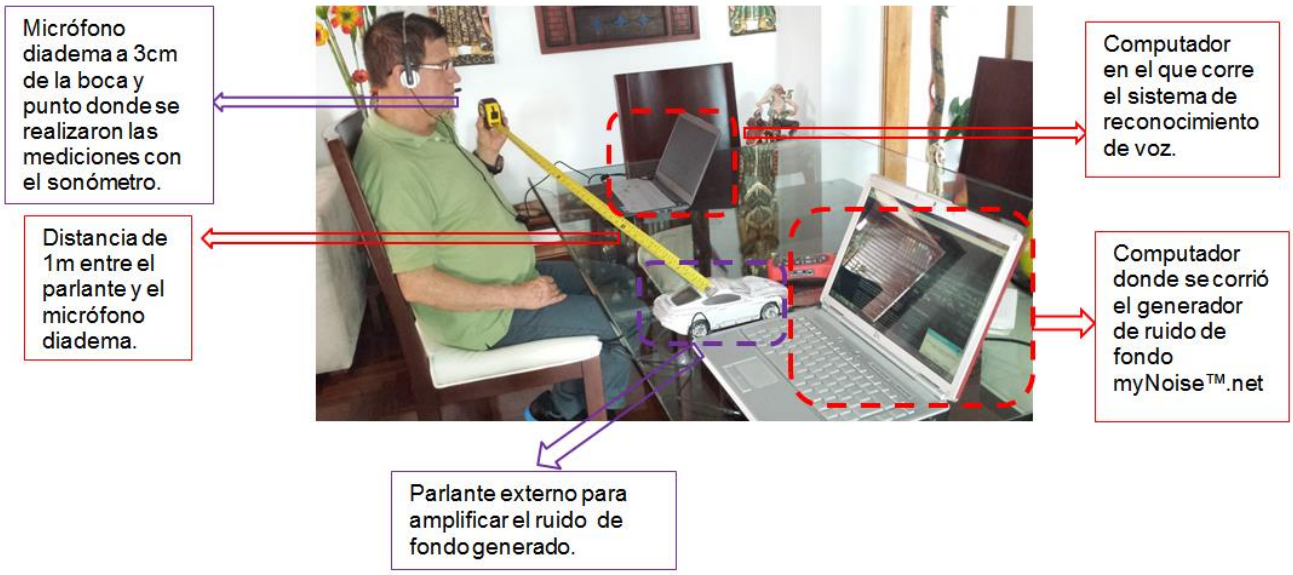

Figura 35: Entorno de las pruebas.

- Datos a obtener de la prueba:

El número de comandos que cada persona debe pronunciar por cada una de las tres pruebas es de 35 comandos, repitiendo cada uno de los mismos 4 veces. De tal manera que por persona se pronuncian en total 140 comandos en cada prueba.

Con las pruebas se desean obtener los siguientes resultados de interés sobre el comportamiento de la aplicación en el reconocimiento de la voz:

- Variaciones en la respuesta de la aplicación de reconocimiento para interlocutores de género masculino en cada uno de los tres rangos de ruido en el entorno.
- Variaciones en la respuesta de la aplicación de reconocimiento para interlocutores de género femenino en cada uno de los tres rangos de ruido en el entorno.
- Diferencias en la respuesta de la aplicación de reconocimiento entre hombres y mujeres para los mismos rangos de nivel de ruido en el entorno.
- Comportamiento general de la aplicación de reconocimiento en cada uno de los tres rangos de ruido en el entorno.

El análisis de los resultados obtenidos en las pruebas se realiza por medio de una matriz de confusión para cada variable a analizar (sexo y nivel de ruido). Dicha matriz es una herramienta estadística de visualización que permite evaluar la eficiencia del sistema de reconocimiento. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases, brindando así una forma para analizar cuales comandos tuvieron errores en su reconocimiento y como se presentaron.

Las filas de la matriz de confusión presentan el listado de los comandos del vocabulario del reconocedor que serán pronunciadas y las columnas contienen los comandos que fueron reconocidos por el sistema. La diagonal de la matriz, presenta el número de comandos que fueron reconocidos correctamente. Cualquier diferencia entre fila y columna representa un error entre el comando reconocido (columna) y el comando del vocabulario (fila).

Una matriz de confusión de 2*2 presenta la estructura que se muestra en la Tabla 4.

	Salidas	
Entradas	(VP) Verdaderos Positivos	(FN) Falsos Negativos
	(FP) Falsos Positivos	(VN) Verdaderos Negativos

Tabla 4: Estructura de una matriz de confusión.

Los parámetros de eficiencia que se van a calcular sobre las matrices de confusión para validar su funcionamiento son:

- **Exactitud**: Definida como la proporción del total de número de predicciones que fueron detectadas correctamente. Se calcula dividiendo el número total de comandos correctamente clasificados por el número total de comandos de referencia.

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN} \quad (16)$$

- **Sensibilidad**: Es la tasa de verdaderos positivos. Mide cuánta información relevante ha extraído el clasificador, expresada como el porcentaje de casos correctamente clasificados en una categoría respecto al total de casos que realmente pertenecen a esa categoría. La

ecuación **(17)** muestra el cálculo de la sensibilidad de una categoría, donde FN (falso negativo) representa el número de casos de esa categoría que reciben, incorrectamente, una etiqueta categórica distinta.

$$Sensibilidad = \frac{VP}{VP + FN} \quad (17)$$

- **Especificidad**: es la tasa de verdaderos negativos. Mide la proporción de negativos que se identificaron correctamente como tal. Se refiere por lo tanto a la capacidad de la prueba para excluir correctamente una condición.

$$Especificidad = \frac{VN}{FP + VN} \quad (18)$$

- **Precisión (Valor predictivo positivo)**: mide qué cantidad de información de la que devuelve el clasificador es correcta, expresada como el porcentaje de casos de una categoría concreta que son correctamente clasificados respecto al total de casos que reciben (correctamente o no) esa misma etiqueta categórica. La ecuación (**19**) muestra el cálculo de la precisión de una categoría, donde VP (verdaderos positivos) representa los casos de esa categoría correctamente clasificados y FP (falsos positivos) representa los casos de otras categorías que, incorrectamente, reciben esa etiqueta categórica. [65]

$$Precisión = \frac{VP}{VP + FP} \quad (19)$$

- **Medida F1**: es la media armónica de la precisión y la sensibilidad; intenta caracterizar el rendimiento de un clasificador mediante un único valor.

$$F1 = \frac{2 * precisión * sensibilidad}{precisión + sensibilidad} \quad (20)$$

Al incluir los valores de precisión y sensibilidad en la ecuación (**20**) se obtiene la siguiente ecuación equivalente

$$F1 = \frac{2VP}{2VP + FP + FN} \quad (21)$$

El instrumento para la recolección de las pruebas por usuario y los comandos por cada gramática se muestra en Anexo 1.

4.2 Resultados y discusión

De las pruebas se obtuvieron nueve matrices de confusión a analizar, tres matrices pertenecientes al género femenino y tres matrices pertenecientes al género masculino por cada nivel de ruido como se establece en la Tabla 3, así como tres matrices con la respuesta general del sistema por el total de las 20 personas que realizaron cada prueba en los niveles de ruido establecido.

Cada una de las tres matrices de confusión diferenciadas por género quedaron conformada por un total de 40 repeticiones por cada comando, puesto que cada una de las 10 personas pronunció el mismo comando 4 veces por prueba, obteniendo así un total de 1400 comandos pronunciados en cada una de las matrices mencionadas.

A cada usuario antes de empezar la prueba se le hacía la aclaración que debía pronunciar los comandos de la misma manera para los tres rangos de nivel de ruido, sin subir la voz en las pruebas dos y tres donde el ruido era mayor (efecto Lombard), debían por lo tanto pronunciar igual de bajo que lo harían en la primera prueba donde se tiene un espacio en silencio Tabla 3. La importancia de esta aclaración se retomará al final de esta sección donde se pudo observar que el sistema desmejora notablemente su respuesta de reconocimiento cuando los comandos se pronuncian con la voz alzada o gritando.

4.2.1 Respuesta del sistema para la prueba #1, rango de nivel de ruido de 35 dB(A) hasta 55 dB(A) para el género masculino y femenino:

Como se observa en la Tabla 5 y Tabla 6, los resultados en un entorno en silencio para ambos géneros tuvieron el mismo comportamiento, obteniéndose un reconocimiento exitoso del 100% de los comandos pronunciados, sin presentarse casos de omisión o de sustitución entre los mismos.

Por lo anterior, los parámetros de eficiencia calculados sobre las matrices de confusión presentadas en la Tabla 5 y Tabla 6 y siguiendo las ecuaciones número (**16**) a la (**21**), presentándose en porcentaje son:

- Exactitud del 100%.
- Sensibilidad en todos los comandos del 100%.
- Especificidad en todos los comandos del 100%.
- Precisión (Valor predictivo positivo) en todos los comandos del 100%.
- Medida F1 en todos los comandos del 100%.

Prueba #1
Mujeres **Rango de 35 dB(A) hasta 55 dB(A)**

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	**Total Comandos**
Mover adelante	40																																		40
Mover atrás		40																																	40
Mover izquierda			40																																40
Mover derecha				40																															40
Mover parar					40																														40
Mover lento						40																													40
Mover rápido							40																												40
Prender luz pasillo								40																											40
Prender luz cocina									40																										40
Apagar luz pasillo										40																									40
Apagar luz alcoba											40																								40
Abrir puerta entrada												40																							40
Cerrar cortina alcoba													40																						40
Abrir cortina entrada														40																					40
Cerrar puerta alcoba															40																				40
Capturar electro																40																			40
Capturar temperatura																	40																		40
Medir electro																		40																	40
Medir oxígeno																			40																40
Tomar presión																				40															40
Tomar pulso																					40														40
ver programas																						40													40
ver silla																							40												40
Ver inicio																								40											40
Ver domótica																									40										40
Ver salud																										40									40
ver configuración																											40								40
Voz detener																												40							40
voz activar																													40						40
Correo médico																														40					40
Correo vecino																															40				40
Correo hermano																																40			40
Aplicación calculadora																																	40		40
Aplicación fecha																																		40	40
Aplicación notas																																		40	40

Tabla 5: Matriz de Confusión para la prueba de mujeres en el rango de nivel de ruido entre 35 dB(A) hasta 55 dB(A)

Prueba #1
Hombres **Rango de 35 dB(A) hasta 55 dB(A)**

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	**Total Comandos**
Mover adelante	40																																		40
Mover atrás		40																																	40
Mover izquierda			40																																40
Mover derecha				40																															40
Mover parar					40																														40
Mover lento						40																													40
Mover rápido							40																												40
Prender luz pasillo								40																											40
Prender luz cocina									40																										40
Apagar luz pasillo										40																									40
Apagar luz alcoba											40																								40
Abrir puerta entrada												40																							40
Cerrar cortina alcoba													40																						40
Abrir cortina entrada														40																					40
Cerrar puerta alcoba															40																				40
Capturar electro																40																			40
Capturar temperatura																	40																		40
Medir electro																		40																	40
Medir oxígeno																			40																40
Tomar presión																				40															40
Tomar pulso																					40														40
ver programas																						40													40
ver silla																							40												40
Ver inicio																								40											40
Ver domótica																									40										40
Ver salud																										40									40
ver configuración																											40								40
Voz detener																												40							40
voz activar																													40						40
Correo médico																														40					40
Correo vecino																															40				40
Correo hermano																																40			40
Aplicación calculadora																																	40		40
Aplicación fecha																																		40	40
Aplicación notas																																		40	40

Tabla 6: Matriz de Confusión para la prueba de hombres en el rango de nivel de ruido entre 35 dB(A) hasta 55 dB(A)

4.2.2 Respuesta del sistema para la prueba #2, rango de nivel de ruido de 60 dB(A) hasta 72 dB(A) para el género masculino y femenino:

La Tabla 7 y Tabla 8 muestran la matriz de confusión para la prueba en mujeres y hombres respectivamente en un entorno donde el nivel de ruido se controló para que permaneciera entre los 60 dB(A) hasta los 72 dB(A) medidos como ya describió con un sonómetro cerca del micrófono diadema que el usuario tenía puesto.

Los resultados para ambos géneros tuvieron un comportamiento casi igual, donde solo se presentó error con uno de los comandos pronunciados. Para el caso de las mujeres, el comando “ver configuración” fue reconocido en una oportunidad como “Tomar presión”, por lo que se dio un error de sustitución. Por su parte, para el caso de los hombres, el comando “Ver inicio” no fue reconocido en una oportunidad, obteniéndose un error de omisión.

Prueba #2
Mujeres **Rango 60 dB(A) hasta 72 dB(A)**

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	**Total Comandos**
Mover adelante	40																																		40
Mover atrás		40																																	40
Mover izquierda			40																																40
Mover derecha				40																															40
Mover parar					40																														40
Mover lento						40																													40
Mover rápido							40																												40
Prender luz pasillo								40																											40
Prender luz cocina									40																										40
Apagar luz pasillo										40																									40
Apagar luz alcoba											40																								40
Abrir puerta entrada												40																							40
Cerrar cortina alcoba													40																						40
Abrir cortina entrada														40																					40
Cerrar puerta alcoba															40																				40
Capturar electro																40																			40
Capturar temperatura																	40																		40
Medir electro																		40																	40
Medir oxígeno																			40																40
Tomar presión																				40															40
Tomar pulso																					40														40
ver programas																						40													40
ver silla																							40												40
Ver inicio																								40											40
Ver domótica																									40										40
Ver salud																										40									40
ver configuración																				1							39								40
Voz detener																												40							40
voz activar																													40						40
Correo médico																														40					40
Correo vecino																															40				40
Correo hermano																																40			40
Aplicación calculadora																																	40		40
Aplicación fecha																																		40	40
Aplicación notas																																			40

Tabla 7: Matriz de Confusión para la prueba de mujeres en el rango de nivel de ruido entre 60 dB(A) hasta 72 dB(A)

Prueba #2
Hombres **Rango 60 dB(A) hasta 72 dB(A)**

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	No reconocido	**Total Comandos**
Mover adelante	40																																			40
Mover atrás		40																																		40
Mover izquierda			40																																	40
Mover derecha				40																																40
Mover parar					40																															40
Mover lento						40																														40
Mover rápido							40																													40
Prender luz pasillo								40																												40
Prender luz cocina									40																											40
Apagar luz pasillo										40																										40
Apagar luz alcoba											40																									40
Abrir puerta entrada												40																								40
Cerrar cortina alcoba													40																							40
Abrir cortina entrada														40																						40
Cerrar puerta alcoba															40																					40
Capturar electro																40																				40
Capturar temperatura																	40																			40
Medir electro																		40																		40
Medir oxígeno																			40																	40
Tomar presión																				40																40
Tomar pulso																					40															40
ver programas																						40														40
ver silla																							40													40
Ver inicio																								39												1	40
Ver domótica																									40												40
Ver salud																										40											40
ver configuración																											40										40
Voz detener																												40									40
voz activar																													40								40
Correo médico																														40							40
Correo vecino																															40						40
Correo hermano																																40					40
Aplicación calculadora																																	40				40
Aplicación fecha																																		40			40
Aplicación notas																																			40		40

Tabla 8: Matriz de Confusión para la prueba de hombres en el rango de nivel de ruido entre 60 dB(A) hasta 72 dB(A)

Los parámetros de eficiencia calculados sobre las matrices de confusión presentadas en la Tabla 7 y Tabla 8, siguiendo las ecuaciones número (**16**) a la (**21**) y presentándose en porcentaje son:

- Exactitud del 99,93% para el reconocimiento tanto en mujeres como en hombres.
- La sensibilidad en 34 de los 35 comandos pronunciados tanto en mujeres como en hombres fue del 100%. El comando “ver configuración” para las mujeres y “Ver inicio” para los hombres obtuvieron el mismo valor de sensibilidad del 97,5%, presentando cada uno de estos un falso negativo.
- En la prueba con mujeres, la especificidad en 34 de los 35 comandos pronunciados es del 100%; el comando “Tomar presión” obtuvo una especificidad del 99,93% al tomar una vez como falso positivo el comando “Ver configuración”. Por su parte, la especificidad en la prueba con hombres fue para todos los comandos del 100% ya que ninguno obtuvo un falso positivo.

- La precisión (Valor predictivo positivo) en la prueba con mujeres, obtuvo un valor del 100% para 34 de los 35 comandos pronunciados; el comando “Tomar presión” obtuvo una precisión del 97,56% dando como falso positivo en una ocación el comando “Ver configuración”. La precisión en la prueba con hombres fue para todos los comandos del 100% sin obtenersen falsos positivos.
- En la prueba con mujeres, la Medida F1 fue del 100% para 33 de los 35 comandos pronunciados; el comando “Tomar presión” obtuvo una medida F1 de 98,77% y el comando “ver configuración” obtuvo una medida F1 de 98,73%. En la prueba con hombres, la Medida F1 fue del 100% para 34 de los 35 comandos pronunciados; solo el comando “Ver inicio” obtuvo una medida del 98,73%.

  La prueba con hombres no ubico ningún comando como falso positivo pero en lugar de ello obtuvo un error de omisión al no identificar en una ocasión el comando pronunciado “Ver inicio” ni relacionarlo con algún otro comando.

4.2.3 Respuesta del sistema para la prueba #3, rango de nivel de ruido de 73 dB(A) hasta 85 dB(A) para el género masculino y femenino:

La matriz de confusión para las pruebas en mujeres y hombres que se muestra en la Tabla 9 y Tabla 10 respectivamente, pertenecen a un entorno donde el nivel de ruido se controló para que permaneciera entre los 73 dB(A) hasta los 85 dB(A). En estas condiciones se percibe un nivel de ruido alto y se hace difícil establecer una comunicación con una persona que este al lado, teniendo que subir la voz para ser escuchado, sin embargo, como ya se mencionó, en la prueba se le advirtió a cada usuario que conservara el mismo tono con que pronunció en el ambiente silencioso, evitando en todo momento subir la voz.

En la prueba con mujeres (Tabla 9), 28 de los 35 comandos pronunciados obtuvieron un reconocimiento exitoso del 100%, los 7 comandos restantes obtuvieron solo errores de omisión al no ser identificados ni reconocidos como otro comando, sin presentarse por lo tanto falsos positivos. Por su parte, en la prueba con hombres (Tabla 9), 21 de los 35 comandos pronunciados obtuvieron un reconocimiento exitoso del 100%, los 14 comandos restantes obtuvieron errores ya sea de omisión o de sustitución.

Prueba #3
Mujeres **Rango 73 dB(A) hasta 85 dB(A)**

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	No reconocido	**Total Comandos**
Mover adelante	40																																			40
Mover atrás		40																																		40
Mover izquierda			40																																	40
Mover derecha				39																															1	40
Mover parar					40																															40
Mover lento						40																														40
Mover rápido							39																												1	40
Prender luz pasillo								40																												40
Prender luz cocina									40																											40
Apagar luz pasillo										40																										40
Apagar luz alcoba											38																								2	40
Abrir puerta entrada												40																								40
Cerrar cortina alcoba													39																						1	40
Abrir cortina entrada														40																						40
Cerrar puerta alcoba															40																					40
Capturar electro																40																				40
Capturar temperatura																	40																			40
Medir electro																		40																		40
Medir oxígeno																			40																	40
Tomar presión																				39															1	40
Tomar pulso																					38														2	40
ver programas																						40														40
ver silla																							40													40
Ver inicio																								39												1	40
Ver domótica																									40												40
Ver salud																										40											40
ver configuración																											40										40
Voz detener																												40									40
voz activar																													40								40
Correo médico																														40							40
Correo vecino																															40						40
Correo hermano																																40					40
Aplicación calculadora																																	40				40
Aplicación fecha																																		40			40
Aplicación notas																																			40		40

Tabla 9: Matriz de Confusión para la prueba de mujeres en el rango de nivel de ruido entre 73 dB(A) hasta 85 dB(A)

Prueba #3
Hombres **Rango 73 dB(A) hasta 85 dB(A)**

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	No reconocido	**Total Comandos**
Mover adelante	40																																			40
Mover atrás		40																																		40
Mover izquierda			39				1																													40
Mover derecha				39																															1	40
Mover parar					40																															40
Mover lento						37												1																	2	40
Mover rápido							40																													40
Prender luz pasillo								39																											1	40
Prender luz cocina									38										1																1	40
Apagar luz pasillo										38																									2	40
Apagar luz alcoba											38																								2	40
Abrir puerta entrada												39																							1	40
Cerrar cortina alcoba													40																							40
Abrir cortina entrada														40																						40
Cerrar puerta alcoba															37																				3	40
Capturar electro																40																				40
Capturar temperatura																	37																		3	40
Medir electro																		40																		40
Medir oxígeno																			40																	40
Tomar presión																				40																40
Tomar pulso																					40															40
ver programas																						40														40
ver silla																							40													40
Ver inicio																		1						38											1	40
Ver domótica																									39										1	40
Ver salud																										39									1	40
ver configuración																											40									40
Voz detener																												40								40
voz activar																													40							40
Correo médico																														39					1	40
Correo vecino																															40					40
Correo hermano																																40				40
Aplicación calculadora																																	40			40
Aplicación fecha																																		40		40
Aplicación notas																																			40	40

Tabla 10: Matriz de Confusión para la prueba de hombres en el rango de nivel de ruido entre 73 dB(A) hasta 85 dB(A)

Los parámetros de eficiencia calculados sobre las matrices de confusión presentadas en la Tabla 9 y Tabla 10, siguiendo las ecuaciones número (**16**) a la (**21**) y presentándose en porcentaje son:

- Exactitud de 99,36% para la prueba en mujeres y de 98,29% para la prueba en hombres.
- La sensibilidad en 28 de los 35 comandos pronunciados en la prueba con mujeres obtuvo un valor del 100%; los otros 7 comandos obtuvieron una sensibilidad superior o igual al 95%. Para la prueba en hombres, la sensibilidad en 21 de los 35 comandos pronunciados obtuvo un valor del 100%; los restantes 14 comandos presentaron una sensibilidad del 92,5% en adelante. La Tabla 11 muestra los resultados de sensibilidad para todos los comandos pronunciados en el rango de nivel de ruido de 73 dB(A) hasta 85 dB(A).
- La especificidad para todos los comandos en la prueba con mujeres es del 100% ya que ninguno obtuvo un falso positivo. Para la prueba con hombres, la especificidad en 32 de los 35

comandos pronunciados obtuvo un valor del 100%; de los restantes, los comandos “Mover rápido y “Medir oxígeno” obtuvieron una especificidad de 99,93% y el comando “Medir electro” obtuvo una especificidad de 99,85%.

- La precisión (Valor predictivo positivo) para todos los comandos en la prueba con mujeres es del 100% al no obtenersen falsos positivos. Para la prueba con hombres, la precisión en 32 de los 35 comandos pronunciados obtuvo un valor del 100%; los comandos “Mover rápido y “Medir oxígeno” obtuvieron una precisión de 97,56% y el comando “Medir electro” obtuvo una precisión de 95,24%.
- La medida F1 en 28 de los 35 comandos pronunciados en la prueba con mujeres obtuvo un valor del 100%; los otros 7 comandos obtuvieron una medida F1 superior o igual al 97,44%. Para la prueba con hombres, la medida F1 en 18 de los 35 comandos pronunciados obtuvo un valor del 100%; los restantes 17 comandos obtuvieron una medida F1 entre el 96,1% y el 98,77%. La Tabla 12 muestra los resultados de la medida F1 para todos los comandos pronunciados en el rango de nivel de ruido de 73 dB(A) hasta 85 dB(A).

	Sensibilidad	
Comando	**Hombres**	**Mujeres**
Mover adelante	100%	100%
Mover atrás	100%	100%
Mover izquierda	97,5%	100%
Mover derecha	97,5%	97,5%
Mover parar	100%	100%
Mover lento	92,5%	100%
Mover rápido	100%	97,5%
Prender luz pasillo	97,5%	100%
Prender luz cocina	95%	100%
Apagar luz pasillo	95%	100%
Apagar luz alcoba	95%	95%
Abrir puerta entrada	97,5%	100%
Cerrar cortina alcoba	100%	97,5%
Abrir cortina entrada	100%	100%
Cerrar puerta alcoba	92,5%	100%
Capturar electro	100%	100%
Capturar temperatura	92,5%	100%
Medir electro	100%	100%
Medir oxígeno	100%	100%
Tomar presión	100%	97,5%
Tomar pulso	100%	95%

	Medida F1	
Comando	**Hombres**	**Mujeres**
Mover adelante	100%	100%
Mover atrás	100%	100%
Mover izquierda	98,73%	100%
Mover derecha	98,73%	98,73%
Mover parar	100%	100%
Mover lento	96,1%	100%
Mover rápido	98,77%	98,73%
Prender luz pasillo	98,73%	100%
Prender luz cocina	97,44%	100%
Apagar luz pasillo	97,44%	100%
Apagar luz alcoba	97,44%	97,44%
Abrir puerta entrada	98,73%	100%
Cerrar cortina alcoba	100%	98,73%
Abrir cortina entrada	100%	100%
Cerrar puerta alcoba	96,1%	100%
Capturar electro	100%	100%
Capturar temperatura	96,1%	100%
Medir electro	97,56%	100%
Medir oxígeno	98,77%	100%
Tomar presión	100%	98,73%
Tomar pulso	100%	97,44%

ver programas	100%	100%
ver silla	100%	100%
Ver inicio	95%	97,5%
Ver domótica	97,5%	100%
Ver salud	97,5%	100%
ver configuración	100%	100%
Voz detener	100%	100%
voz activar	100%	100%
Correo médico	97,5%	100%
Correo vecino	100%	100%
Correo hermano	100%	100%
Aplicación calculadora	100%	100%
Aplicación fecha	100%	100%
Aplicación notas	100%	100%

Tabla 11: Resultados de la sensibilidad para el rango de nivel de ruido entre 73 dB(A) hasta 85 dB(A)

ver programas	100%	100%
ver silla	100%	100%
Ver inicio	97,44%	98,73%
Ver domótica	98,73%	100%
Ver salud	98,73%	100%
ver configuración	100%	100%
Voz detener	100%	100%
voz activar	100%	100%
Correo médico	98,73%	100%
Correo vecino	100%	100%
Correo hermano	100%	100%
Aplicación calculadora	100%	100%
Aplicación fecha	100%	100%
Aplicación notas	100%	100%

Tabla 12: Resultados de la medida F1 para el rango de nivel de ruido entre 73 dB(A) hasta 85 dB(A)

La Tabla 13 muestra un resumen comparativo de los parámetros de eficiencia calculados sobre las matrices de confusión para los tres ambientes de prueba en hombres y en mujeres. Se puede analizar de dichos resultados que el sistema de reconocimiento de voz en español para un vocabulario cerrado e independiente del hablante, no presenta diferencias significativas en su desempeño al responder ante hombres y mujeres.

En la prueba #1donde el rango de nivel de ruido se mantuvo entre 35 dB(A) hasta 55 dB(A), todos los parámetros de eficiencia de la matriz de confusión obtuvieron una respuesta del 100% tanto en hombres como en mujeres.

En la prueba #2 donde el rango de nivel de ruido se mantuvo entre 60 dB(A) hasta 72 dB(A), 34 de los 35 comandos fueron reconocidos exitosamente en todas las oportunidades para ambos géneros, en el caso de las mujeres el error presente se dio por sustitución y en de los hombres se dio por omisión.

Para la prueba #3 donde el rango de nivel de ruido se mantuvo entre 73 dB(A) hasta 85 dB(A), la respuesta aunque sigue siendo muy pareja, presenta en el caso de las mujeres valores levemente superiores en todos los parámetros de eficiencia que los obtenidos por los hombres; en esta prueba la sensibilidad que muestra el porcentaje de casos correctamente clasificados en una categoría respecto al total de casos que realmente pertenecen a esa categoría, es el parámetro que presenta el menor valor al compararlo con los demás parámetros aunque sigue teniendo una muy buena

respuesta que para el caso de los hombres cuyo valor fue el mas bajo es de 92,5%; por su parte, en esta prueba, el parámetro que presentó el valor mas alto es la especificidad, el cual mide la proporción de negativos que se identificaron correctamente como tal, presentándose en la respuesta mas baja que de igual manera fue en la prueba con hombres un valor de 99,85%.

	Prueba #1 **35 dB(A) hasta 55 dB(A)**		**Prueba #2** **60 dB(A) hasta 72 dB(A)**		**Prueba #3** **73 dB(A) hasta 85 dB(A)**	
	Mujeres	**Hombres**	**Mujeres**	**Hombres**	**Mujeres**	**Hombres**
Exactitud	100%	100%	99,93%	99,93%	99,36%	98,29%
Sensibilidad	100% en todos los comandos	100% en todos los comandos	97,5% en solo uno de los comandos. El resto 100%	97,5% en solo uno de los comandos. El resto 100%	Superior al 94,99% en siete de los comandos. El resto 100%.	Superior al 92,4% en catorce de los comandos. El resto 100%.
Especificidad	100% en todos los comandos	100% en todos los comandos	99,93% en solo uno de los comandos. El resto 100%	100% en todos los comandos	100% en todos los comandos	Superior al 99,8% en tres de los comandos. El resto 100%.
Precisión	100% en todos los comandos	100% en todos los comandos	97,56% en solo uno de los comandos. El resto 100%	100% en todos los comandos	100% en todos los comandos	Superior al 95,2% en tres de los comandos. El resto 100%.
Medida F1	100% en todos los comandos	100% en todos los comandos	Superior al 98,7% en dos de los comandos. El resto 100%	98,73% en solo uno de los comandos. El resto 100%	Superior al 97,4% en siete de los comandos. El resto 100%.	Superior al 96% en diez y siete de los comandos. El resto 100%.

Tabla 13: Resumen de los parámetros de eficiencia calculados sobre las matrices de confusión para los tres ambientes de prueba en hombres y en mujeres.

Observando la Tabla 13 se pueden comparar también los resultados entre las pruebas con mujeres en cada uno de los tres entornos así como comparar separadamente los resultados entre las pruebas con hombres. Se observa que a medida que el nivel de ruido se incrementa, la respuesta del sistema de reconocimiento va disminuyendo, permaneciendo casi igual la respuesta para las pruebas #1 y #2 donde la poca disminución en los valores de los parámetros de eficiencia se debió para ambos géneros por un solo comando que en una oportunidad no fue correctamente clasificado.

Para el caso de las mujeres, el valor de exactitud que indica la proporción del total de número de predicciones que fueron detectadas correctamente cambió del 100% para un ambiente en silencio al 99,36% para el entorno mas ruidoso (prueba #3), siendo esta diferencia de apenas 0,64%. De manera similar, para el caso de los hombres, el valor de exactitud cambió del 100% para un ambiente

en silencio al 98,29% para el entorno mas ruidoso, siendo la diferencia de 1,71%, valor levemente superior al presentado en la mujeres.

Una observación presentada en la prueba #3 donde la proporción de comandos no reconocidos exitosamente disminuyo tiene que ver con el nivel de confidencia establecido en el programa para las pruebas. Para todas las personas participantes en la prueba se fijó el mismo nivel de confidencia, sin embargo, con la ayuda del monitoreo realizado a través de los datos que presenta la pestaña "Pruebas" en el programa, Figura 33, se observó que la aplicación si reconocía efectivamente muchos de los comandos pero con un nivel de confidencia menor que el establecido, por lo cual no alcanzaba a pasar como un comando válido reconocido.

4.2.4 Respuesta general del sistema para las pruebas en los tres rango de nivel de ruido (sin discriminar por género):

Adicional al análisis realizado en los puntos anteriores donde se discriminó el comportamiento del sistema según el género, se elaboró también una matriz de confusión por cada uno de los tres rangos de nivel de ruido para el total de las 20 personas que participaron en la prueba (total de hombres y mujeres), extrayendo de allí el comportamiento general del sistema de reconocimiento diferenciado solo por nivel de ruido. Como cada una de los 20 participantes repitió el mismo comando 4 veces, el total de veces que se pronunció cada comando es de 80, lo cual para los 35 comandos de la prueba da un total de 2800 comandos pronunciados.

La Tabla 14 muestra la matriz de confusión para la prueba en un entorno en silencio donde el rango de nivel de ruido se mantuvo entre 35 dB(A) hasta 55 dB(A), en ella se aprecia que se obtuvo un reconocimiento exitoso del 100% de los comandos pronunciados, sin presentarse casos de omisión o de sustitución entre los mismos. Los parámetros de eficiencia calculados sobre esta matriz, siguiendo las ecuaciones número (**16**) a la (**21**) y presentándose en porcentaje son:

- Exactitud del 100%.
- Sensibilidad en todos los comandos del 100%.
- Especificidad en todos los comandos del 100%.
- Precisión (Valor predictivo positivo) en todos los comandos del 100%.
- Medida F1 en todos los comandos del 100%.

Prueba #1: Rango de 35 dB(A) hasta 55 dB(A)

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	**Total Comandos**
Mover adelante	80																																		80
Mover atrás		80																																	80
Mover izquierda			80																																80
Mover derecha				80																															80
Mover parar					80																														80
Mover lento						80																													80
Mover rápido							80																												80
Prender luz pasillo								80																											80
Prender luz cocina									80																										80
Apagar luz pasillo										80																									80
Apagar luz alcoba											80																								80
Abrir puerta entrada												80																							80
Cerrar cortina alcoba													80																						80
Abrir cortina entrada														80																					80
Cerrar puerta alcoba															80																				80
Capturar electro																80																			80
Capturar temperatura																	80																		80
Medir electro																		80																	80
Medir oxígeno																			80																80
Tomar presión																				80															80
Tomar pulso																					80														80
ver programas																						80													80
ver silla																							80												80
Ver inicio																								80											80
Ver domótica																									80										80
Ver salud																										80									80
ver configuración																											80								80
Voz detener																												80							80
voz activar																													80						80
Correo médico																														80					80
Correo vecino																															80				80
Correo hermano																																80			80
Aplicación calculadora																																	80		80
Aplicación fecha																																		80	80
Aplicación notas																																			80

Tabla 14: Matriz de Confusión para el rango de nivel de ruido entre 35 dB(A) hasta 55 dB(A) sin discriminar género del locutor.

La Tabla 15 por su parte, muestra la matriz de confusión para la prueba en un entorno donde el nivel de ruido se controló para que permaneciera entre 60 dB(A) hasta 72 dB(A), en este caso se observa un error de sustitución para el comando “ver configuración” y un error de omisión para el comando “Ver inicio”. Los parámetros de eficiencia calculados sobre esta matriz, siguiendo las ecuaciones número (**16**) a la (**21**) y presentándose en porcentaje son:

- Exactitud del 99,93% para el reconocimiento general del sistema.
- La sensibilidad en 33 de los 35 comandos pronunciados fue del 100%. Los comandos “ver configuración” y “Ver inicio” obtuvieron el mismo valor de sensibilidad del 98,75% presentando cada uno de estos un falso negativo.
- La especificidad en 34 de los 35 comandos pronunciados es del 100%; el comando “Tomar presión” obtuvo una especificidad del 99,96% al tomar una vez como falso positivo el comando “Ver configuración”.

- La precisión (Valor predictivo positivo) obtuvo un valor del 100% para 34 de los 35 comandos pronunciados; el comando “Tomar presión” obtuvo una precisión del 98,77% dando como falso positivo en una ocación el comando “Ver configuración”.
- La Medida F1 fue del 100% para 32 de los 35 comandos pronunciados; el comando “Tomar presión” obtuvo una medida F1 de 99,38% , los comandos “Ver inicio” y “ver configuración” obtuvieron una medida F1 de 99,37%.

Prueba #2: Rango 60 dB(A) hasta 72 dB(A)

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	No reconocido	Total Comandos
Mover adelante	80																																			80
Mover atrás		80																																		80
Mover izquierda			80																																	80
Mover derecha				80																																80
Mover parar					80																															80
Mover lento						80																														80
Mover rápido							80																													80
Prender luz pasillo								80																												80
Prender luz cocina									80																											80
Apagar luz pasillo										80																										80
Apagar luz alcoba											80																									80
Abrir puerta entrada												80																								80
Cerrar cortina alcoba													80																							80
Abrir cortina entrada														80																						80
Cerrar puerta alcoba															80																					80
Capturar electro																80																				80
Capturar temperatura																	80																			80
Medir electro																		80																		80
Medir oxígeno																			80																	80
Tomar presión																				80																80
Tomar pulso																					80															80
ver programas																						80														80
ver silla																							80													80
Ver inicio																								79											1	80
Ver domótica																									80											80
Ver salud																										80										80
ver configuración																				1							79									80
Voz detener																												80								80
voz activar																													80							80
Correo médico																														80						80
Correo vecino																															80					80
Correo hermano																																80				80
Aplicación calculadora																																	80			80
Aplicación fecha																																		80		80
Aplicación notas																																			80	80

Tabla 15: Matriz de Confusión para el rango de nivel de ruido entre 60 dB(A) hasta 72 dB(A) sin discriminar género del locutor.

Los resultados de la prueba #3 donde el nivel de ruido se controló para que permaneciera entre 73 dB(A) hasta 85 dB(A) se observan en la Tabla 16. Los parámetros de eficiencia calculados sobre esta matriz, siguiendo las ecuaciones número (**16**) a la (**21**) y presentándose en porcentaje son:

- Exactitud de 98,82% para el reconocimiento general del sistema.

- La sensibilidad en 17 de los 35 comandos pronunciados obtuvo un valor del 100%; para los restantes comandos, la sensibilidad obtuvo valores superiores o igual al 95%. Los resultados de sensibilidad para todos los comandos pronunciados en el rango de nivel de ruido de 73 dB(A) hasta 85 dB(A) se muestran en Tabla 17.
- La especificidad en 32 de los 35 comandos pronunciados obtuvo un valor del 100%; de los restantes, los comandos “Mover rápido y “Medir oxígeno” obtuvieron una especificidad de 99,96% y el comando “Medir electro” obtuvo una especificidad de 99,93%.
- La precisión (Valor predictivo positivo) en 32 de los 35 comandos pronunciados obtuvo un valor del 100%; el comando “Mover rápido” obtuvo una precisión de 98,75%, el comando “Medir oxígeno” obtuvo un valor de 98,77% y el comando “Medir electro” obtuvo una precisión de 97,56%.
- La medida F1 en 15 de los 35 comandos pronunciados obtuvo un valor del 100%; los restantes comandos obtuvieron una medida F1 entre el 97,44% y el 99,37%. La Tabla 18 muestra los resultados de la medida F1 para todos los comandos pronunciados en el rango de nivel de ruido de 73 dB(A) hasta 85 dB(A).

Prueba #3: Rango 73 dB(A) hasta 85 dB(A)

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	Capturar electro	Capturar temperatura	Medir electro	Medir oxígeno	Tomar presión	Tomar pulso	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	No reconocido	**Total Comandos**
Mover adelante	80																																			80
Mover atrás		80																																		80
Mover izquierda			79				1																													80
Mover derecha				78																															2	80
Mover parar					80																															80
Mover lento						77												1																	2	80
Mover rápido							79																												1	80
Prender luz pasillo								79																											1	80
Prender luz cocina									78										1																1	80
Apagar luz pasillo										78																									2	80
Apagar luz alcoba											76																								4	80
Abrir puerta entrada												79																							1	80
Cerrar cortina alcoba													79																						1	80
Abrir cortina entrada														80																						80
Cerrar puerta alcoba															77																				3	80
Capturar electro																80																				80
Capturar temperatura																	77																		3	80
Medir electro																		80																		80
Medir oxígeno																			80																	80
Tomar presión																				79															1	80
Tomar pulso																					78														2	80
ver programas																						80														80
ver silla																							80													80
Ver inicio																		1						77											2	80
Ver domótica																									79										1	80
Ver salud																										79									1	80
ver configuración																											80									80
Voz detener																												80								80
voz activar																													80							80
Correo médico																														79					1	80
Correo vecino																															80					80
Correo hermano																																80				80
Aplicación calculadora																																	80			80
Aplicación fecha																																		80		80
Aplicación notas																																			80	80

Tabla 16: Matriz de Confusión para el rango de nivel de ruido entre 73 dB(A) hasta 85 dB(A) sin discriminar género del locutor.

Sensibilidad (sin discriminar género)	
Comando	**Rango de 73 dB(A) hasta 85 dB(A)**
Mover adelante	100%
Mover atrás	100%
Mover izquierda	98,75%
Mover derecha	97,5%
Mover parar	100%
Mover lento	96,25%
Mover rápido	98,75%
Prender luz pasillo	98,75%
Prender luz cocina	97,5%
Apagar luz pasillo	97,5%
Apagar luz alcoba	95%
Abrir puerta entrada	98,75%

Medida F1 (sin discriminar género)	
Comando	**Rango de 73 dB(A) hasta 85 dB(A)**
Mover adelante	100%
Mover atrás	100%
Mover izquierda	99,37%
Mover derecha	98,73%
Mover parar	100%
Mover lento	98,09%
Mover rápido	98,75%
Prender luz pasillo	99,37%
Prender luz cocina	98,73%
Apagar luz pasillo	98,73%
Apagar luz alcoba	97,44%
Abrir puerta entrada	99,37%

Cerrar cortina alcoba	98,75%
Abrir cortina entrada	100%
Cerrar puerta alcoba	96,25%
Capturar electro	100%
Capturar temperatura	96,25%
Medir electro	100%
Medir oxígeno	100%
Tomar presión	98,75%
Tomar pulso	97,5%
ver programas	100%
ver silla	100%
Ver inicio	96,25%
Ver domótica	98,75%
Ver salud	98,75%
ver configuración	100%
Voz detener	100%
voz activar	100%
Correo médico	98,75%
Correo vecino	100%
Correo hermano	100%
Aplicación calculadora	100%
Aplicación fecha	100%
Aplicación notas	100%

Tabla 17: Resultados de la sensibilidad para el rango de nivel de ruido entre 73 dB(A) hasta 85 dB(A), sin discriminar género.

Cerrar cortina alcoba	99,37%
Abrir cortina entrada	100%
Cerrar puerta alcoba	98,09%
Capturar electro	100%
Capturar temperatura	98,09%
Medir electro	98,77%
Medir oxígeno	99,38%
Tomar presión	99,37%
Tomar pulso	98,73%
ver programas	100%
ver silla	100%
Ver inicio	98,09%
Ver domótica	99,37%
Ver salud	99,37%
ver configuración	100%
Voz detener	100%
voz activar	100%
Correo médico	99,37%
Correo vecino	100%
Correo hermano	100%
Aplicación calculadora	100%
Aplicación fecha	100%
Aplicación notas	100%

Tabla 18: Resultados de la medida F1 para el rango de nivel de ruido entre 73 dB(A) hasta 85 dB(A), sin discriminar género.

La Tabla 19 muestra un resumen comparativo de los parámetros de eficiencia calculados sobre las matrices de confusión para los tres ambientes con el total de las 20 personas que participaron en la prueba sin discriminar por género.

	Resultados con el total de los participantes (sin discriminar género)		
	Prueba #1 **35 dB(A) hasta 55 dB(A)**	**Prueba #2** **60 dB(A) hasta 72 dB(A)**	**Prueba #3** **73 dB(A) hasta 85 dB(A)**
Exactitud	100%	99,93%	98,82%
Sensibilidad	100% en todos los comandos	98,75% en solo dos de los comandos. El resto 100%	Superior o igual al 95% en todos los comandos.
Especificidad	100% en todos los comandos	99,96% en solo uno de los comandos. El resto 100%	Superior al 99,9% en tres de los comandos. El resto 100%.
Precisión	100% en todos los comandos	98,77% en solo uno de los comandos. El resto 100%	Superior al 97,5% en tres de los comandos. El resto 100%.

Medida F1	100% en todos los comandos	Superior al 99,3%. en tres de los comandos. El resto 100%	Superior al 97,4% en todos los comandos.

Tabla 19: Resumen de los parámetros de eficiencia calculados sobre las matrices de confusión para los tres ambientes de prueba sin discriminar por género.

En generar el sistema de reconocimiento responde muy bien en los tres ambientes de prueba, dándose una leve desmejora a medida que el ruido en el ambiente aumenta. Sin embargo en el entorno de mayor ruido (prueba #3) el parámetro que obtuvo el valor mas bajo fue la sensibilidad en donde para uno de los comandos dio un resultado del 95%. La exactitud vario del 100% para el ambiente en silencio (prueba #1) a solo el 98,82% en el ambiente con el ruido mas alto (prueba #3). En cuanto a especificidad, precisión y medida F1 el cambio en el comportamiento en los tres rangos de niveles de ruido fue muy poco, presentándose diferencias de no mas del 2,6% entre ellas.

4.2.5 Observación respecto a la influencia de subir la voz al momento de pronunciar los comandos:

Como ya se mencionó en la descripción de las pruebas realizadas, a cada participante se le dio la instrucción de que debía pronunciar los comandos de la misma manera para los tres rangos de nivel de ruido, sin subir la voz en las pruebas dos y tres donde el ruido era mayor. Dicha tendencia involuntaria a incrementar el esfuerzo vocal cuando se habla en un lugar ruidoso con el fin de mejorar la audibilidad de la voz se conoce como *efecto Lombard* e interfiere enormemente en la respuesta del reconocedor ya que los cambios al subir la voz afectan no solo a la sonoridad, sino también a factores como el tono, el rango y la duración del sonido de las sílabas. En la sección 2.5, problemas en la detección, se señaló que cuando un locutor habla en presencia de ruido, han encontrado que el primer formante de una vocal tiende a crecer mientras que el segundo decrece y que la caída espectral decrece en las frecuencias bajas y aumenta en las altas para la mayoría de las vocales [50].

Para comprobar el efecto de subir la voz al momento de pronunciar los comandos en un ambiente ruidoso, se realizó una prueba con los mismos rangos de nivel de ruido de la prueba #3 (entre 73 dB(A) hasta 85 dB(A)) con tres de los participantes que de igual manera debían repetir cada comando 4 veces. La matriz de confusión resultante se puede observar en el Anexo 2.

En esta prueba el valor de exactitud de la matriz de confusión bajo a un 55,77%, cuando en las pruebas anteriores todos los resultados habían sido mayores al 98%.

Solo el 7,7% de los comandos obtuvieron un valor de sensibilidad del 100%, el 35,9% de los comandos obtuvieron un valor inferior al 50% y el 56,4% de los comandos obtuvieron un valor entre el 50% y el 91,7%.

El valor de especificidad fue del 100% para el 79,5% de los comandos; el valor mas bajo presentado de este parámetro para el resto de los comandos fue del 84,2% referente al comando “tomar electro”.

De manera similar a la especificidad, el valor de precisión fue del 100% para el 79,5% de los comandos, los demás comandos presentaron valores de precisión entre 19% (comando “tomar electro”) y el 87,5%.

Con los resultados anteriores se puede observar que el sistema desmejora notablemente su respuesta de reconocimiento al alzar la voz para pronunciar los comandos en el ambiente ruidoso, esfuerzo este innecesario ya que el comando se le está diciendo es a un micrófono que se encuentra cercano a la boca.

4.3 Pruebas Complementarias:

4.3.1. Valor de confidencia adecuado para los comandos de primer nivel:

El valor de confidencia, como se mencionó en la sección 3.5.2, es un parámetro que da una restricción de nivel de confianza al reconocedor; si el valor es muy bajo, puede detectar erróneamente palabras pronunciadas que no están en el vocabulario como válidas, y si es muy alto puede bloquear una mayor cantidad de frases que si son correctas y tomarlas como no válidas. El rango en el que se puede fijar el nivel de confidencia va entre 0 (mínimo) y 1(máximo).

Con el fin de establecer un valor de confidencia adecuado para los comandos fijos que componen el vocabulario cerrado de la aplicación, se realizó una prueba que incluye las 13 clases de primer nivel; de cada una de éstas 13 clases se desprenden las diferentes frases que componen los comandos de la aplicación. La Tabla 20 muestra la palabra que corresponde a cada clase; de forma aleatoria cada una de ellas se pronunció en un total de 20 veces, anotando el valor de confidencia con que eran reconocidas según el monitoreo del comportamiento del reconocedor de voz, visualizado en la pestaña “Pruebas” de la aplicación, como se muestra en la Figura 36.

	Comando de primer nivel
Clase 1	'abrir'
Clase 2	'apagar'
Clase 3	'aplicación'
Clase 4	'capturar'
Clase 5	'cerrar'
Clase 6	'correo'
Clase 7	'enviar'
Clase 8	'medir'
Clase 9	'mover'
Clase 10	'prender'
Clase 11	'tomar'
Clase 12	'ver'
Clase 13	'voz'

Tabla 20: Clases de primer nivel

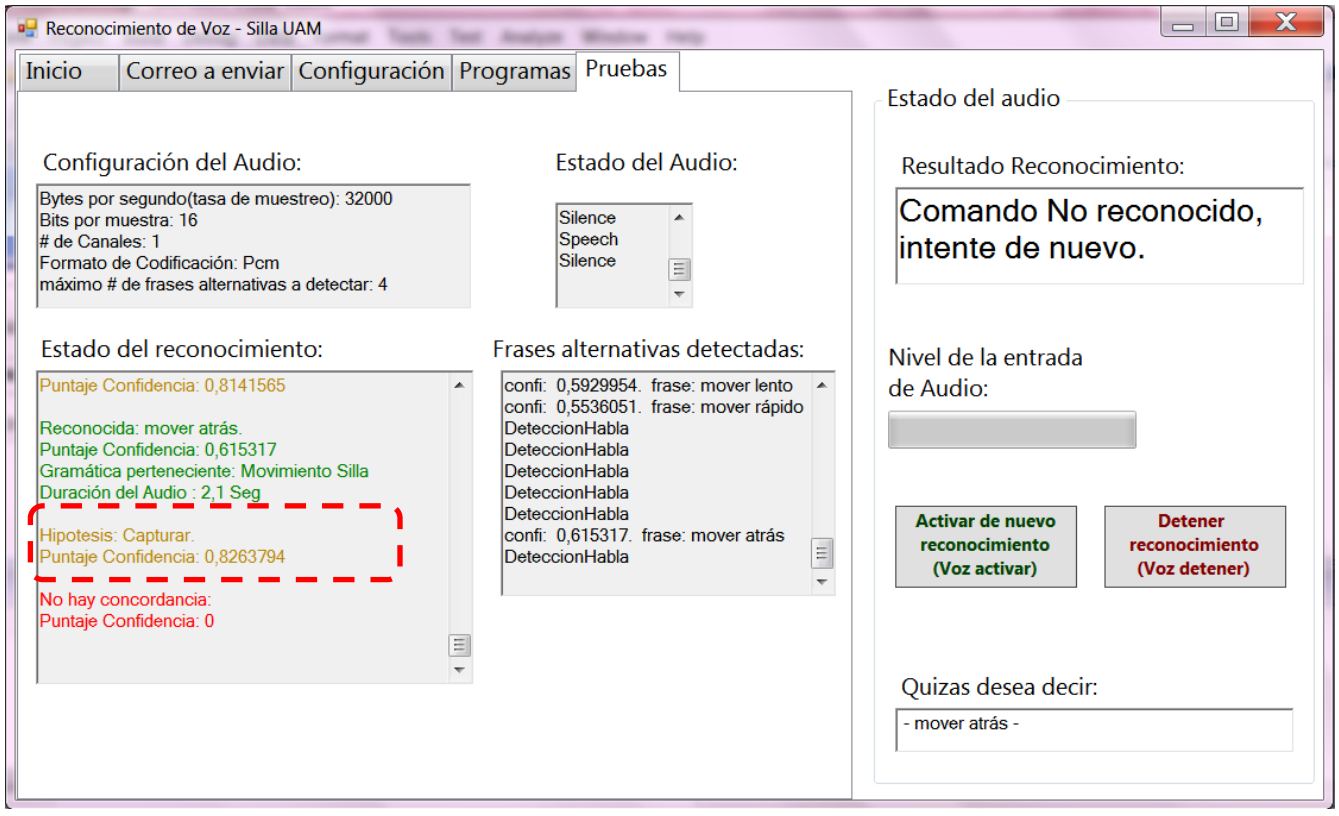

Figura 36: Monitoreo del comportamiento del reconocedor de voz.

La Figura 37 muestra la media del valor de confidencia para cada una de las 13 clases de la Tabla 20. Mientras que la Figura 38 muestra los resultados del valor de confidencia en una gráfica tipo boxplot.

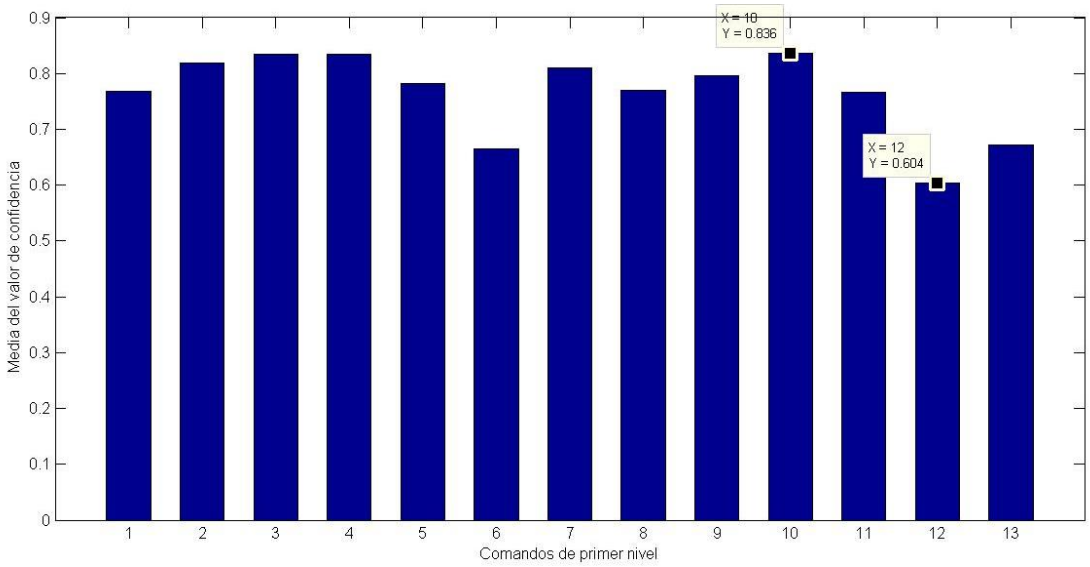

Figura 37: Media del valor de confidencia para las 13 clases de primer nivel

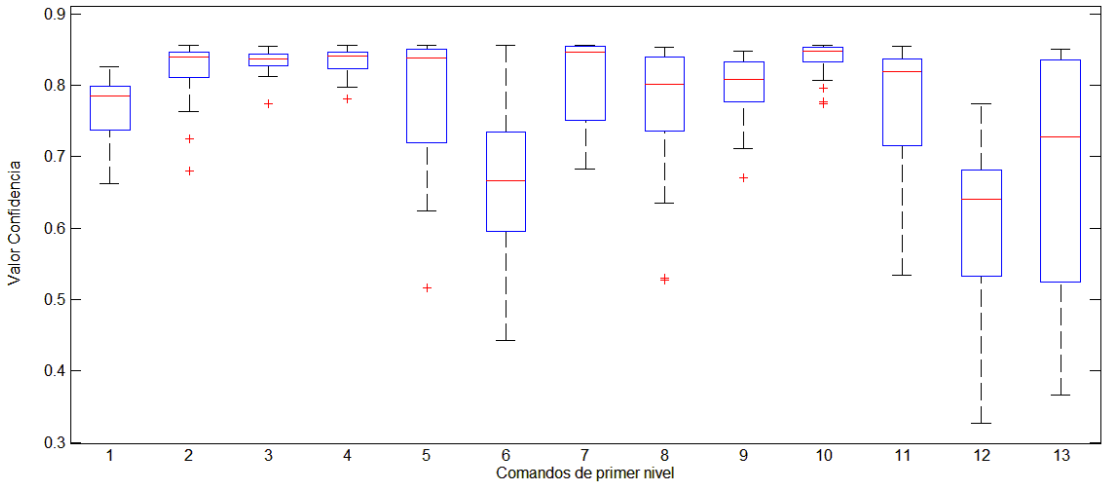

Figura 38: Distribución del valor de confidencia para las 13 clases de primer nivel.

Según Figura 37, el 76.9% de las clases obtuvieron un valor de confidencia con una media por encima de 0.7, la clase 12 (‘ver’) presentó el valor de media más bajo con 0.604, seguido por la clase

6 (‘correo’) con 0.665; por su parte la clase 10 (‘prender’) obtuvo el valor más alto en la media con 0.836, seguido por las clases 3 y 4 (‘aplicación’ y ‘capturar’) con un valor de media de 0.834.

La Figura 38 muestra que de las 20 veces que se pronunció cada palabra, las clases que presentaron menor dispersión en la distribución del valor de confidencia aceptado son la clase 3 (‘aplicación’), clase 10 (‘prender’) y la clase 4 (‘capturar’), con diferencias inferiores a 0.058 entre el menor y el mayor de los valores de confidencia. Por otro lado, las clases con mayor dispersión en la distribución son la clase 13 (‘voz’), clase 12 (‘ver’) y clase 6 (‘correo’), con diferencias de 0.4851, 0.4481 y 0.4145 respectivamente entre el menor y el mayor de los datos. La clase 12 (‘ver’) presentó el valor de confidencia más bajo de todos los casos con 0.3265 y las clases 10, 6 y 5 (‘prender’, ‘correo’ y ‘cerrar’) presentaron el caso con el valor de confidencia más alto, siendo el mismo valor para las tres de 0.8565.

Según resultados anteriores, se puede fijar un valor de confidencia de 0.6 al momento de configurar la restricción de aceptación del reconocedor, valor superado en la media de todas las clases de la Tabla 20; con lo que se pretende que los comandos pronunciados válidos si sean aceptados como tal pero a la vez exista un nivel de rechazo para los casos en los que hay comandos supuestamente reconocidos pero que por su bajo nivel de confianza hay gran incertidumbre de que efectivamente el proceso de reconocimiento este en lo correcto.

Es también importante destacar de los resultados que las palabras ‘voz’, ‘ver’ y ‘correo’ pueden presentar mayores errores al momento de ser reconocidas dada su gran dispersión en la Figura 38, a que el 25% de los datos en cada uno de ellas están por debajo de un valor de confidencia de 0.6 y a que son las palabras que presentaron los casos con valores de confidencia más bajos. Por lo anterior, para futuras mejoras del presente trabajo será recomendado cambiar las tres palabras en mención por comandos que tengan más de una sílaba y que no presenten problemas comunes en la pronunciación como es el caso de la doble r que algunas personas no son capaces de pronunciar.

4.3.2. Respuesta de la aplicación ante la entrada de comandos erróneos:

Se realizó un monitoreo de la respuesta del sistema cuando se pronuncian los comandos predefinidos cambiando la última palabra que compone a los mismos, tanto por palabras fonéticamente parecidas como por palabras fonéticamente diferentes. El valor de confidencia para la prueba se estableció en 0.6. La Tabla 21 y Tabla 22 muestra los resultados de la prueba.

Comando Correcto		Comando fonéticamente parecido (Parónimos)	Respuesta de la aplicación	Comando fonéticamente diferente	Respuesta de la aplicación
Mover	Adelante	Mover caminante	Mover adelante	Mover dinosaurio	No reconocido
		Mover admirante	Mover adelante		
	Derecha	Mover pereza	Mover derecha		
		Mover teresa	Mover derecha		
	Atrás	Mover trazar	mover parar		
		Mover halar	mover parar		
	Izquierda	Mover piedra	mover izquierda		
		Mover histeria	mover izquierda		
	lento	Mover cerdo	mover izquierda		
		Mover pleno	mover lento		
	rápido	Mover calido	mover rápido		
		Mover palido	mover rápido		
Ver	Silla	Ver misa	ver inicio	Ver computador	No reconocido
		Ver pizza	ver inicio		
	Domótica	Ver robótica	ver domótica		
		Ver exótica	ver domótica		
	Salud	Ver mamut	ver domótica		
		Ver shampoo	ver programas		
	Inicio	Ver precipicio	ver inicio		
		Ver mauricio	ver inicio		
	configuración	Ver entonación	No reconocido		
		Ver estimulación	ver configuración		
	Programas	Ver moradas	ver programas		
		Ver sotanas	ver programas		
	Pruebas	Ver ruedas	ver pruebas		
		Ver muelas	No reconocido		
Voz	Activar	Voz caminar	Voz activar	Voz fuerte	No reconocido
		Voz admirar	Voz activar		
	Detener	Voz emerger	voz detener		
		Voz merecer	No reconocido		
Capturar	Electro	Capturar estereo	capturar presión	Capturar botella	No reconocido
		Capturar magneto	capturar electro		
	Presión	Capturar tensión	capturar presión		
		Capturar porción	No reconocido		
	Temperatura	Capturar prematura	No reconocido		
		Capturar soldadura	capturar temperatura		
	Pulso	Capturar curso	capturar pulso		
		Capturar ruso	capturar pulso		
	Oxígeno	Capturar exagono	capturar presión		
		Capturar polimero	capturar pulso		

Tabla 21: Respuesta ante la entrada de comandos erróneos.

Comando Correcto			Comando fonéticamente parecido (Parónimos)	Respuesta de la aplicación	Comando fonéticamente diferente	Respuesta de la aplicación
Prender	Luz	Pasillo	Prender luz camino	Prender luz pasillo	Prender luz comedor	No reconocido
			Prender luz racimo	Prender luz pasillo		
		Cocina	Prender luz piscina	Prender luz cocina		
			Prender luz marina	Prender luz cocina		
		Entrada	Prender luz mesada	Prender luz entrada		
			Prender luz prestada	Prender luz entrada		
		Alcoba	Prender luz anchoa	Prender luz alcoba		
			Prender luz arroba	Prender luz alcoba		
Abrir	Puerta	Entrada	Abrir puerta pesada	Prender puerta entrada	Abrir puerta fotografía	No reconocido
			Abrir puerta mesada	Prender puerta entrada		
		Alcoba	Abrir puerta escoba	Prender puerta alcoba		
			Abrir puerta joroba	Prender puerta alcoba		

Tabla 22: Respuesta ante la entrada de comandos erróneos.

Como se observa en la Tabla 21 y Tabla 22, para el caso de los comandos fonéticamente diferentes el sistema los rechazó adecuadamente dando como resultado ningún comando reconocido; por su parte, para el caso de los comandos fonéticamente parecidos, el sistema los aceptó como una entrada válida en el 90.4% de las veces.

Los resultados anteriores demuestran la importancia del comando de voz y respectivo botón que detiene el reconocimiento en la aplicación desarrollada, Figura 18, ya que si el usuario desea entablar una conversación con alguien o pronunciar palabras que no van destinadas a los comandos por voz, el sistema de reconocimiento se deshabilita para que éste no acepte comandos en dicho momento, hasta que el usuario active de nuevo el reconocimiento por medio de su respectivo botón o comando de voz correspondiente. De igual manera, se hace importante que el usuario de la silla de ruedas se entrene correctamente con cuales son los comandos a pronunciar y se fije en las ayudas visuales que tiene la aplicación para recordar los mismos, ya que si por error confunde una de las palabras que componen el comando, el sistema puede aceptarlo erróneamente como válido y dar como respuesta un comando no deseado de los que componen el vocabulario de la aplicación.

5. CONCLUSIONES

- Se implementó una interfaz gráfica que permite observar la retroalimentación del comportamiento del sistema, informando de manera visual al momento de pronunciar las frases si el comando es reconocido y en caso de serlo muestra cual es el comando, así como emite una retroalimentación auditiva. La interfaz también informa si no se obtiene reconocimiento o si el sistema se encuentra desactivado por el usuario. Así mismo, permitió usar una estructura grafica que facilita la visualización de los comandos a pronunciar. El desplazamiento entre las pestañas del programa se puede realizar por comandos de voz y tiene una función que permite activar o suspender el sistema de reconocimiento (aspecto clave si el usuario desea que el sistema momentáneamente deje de reconocer para entablar por ejemplo una conversación con otra persona o evitar en general que la aplicación reconozca comandos cuando no se les están dictando).

- Si bien al momento de plantearse el proyecto, se definió como uno de los objetivos específicos el desarrollar un modelo de lenguaje para el reconocimiento de comandos de voz en español, se encontró posteriormente tras la investigación de las diferentes herramientas existentes, que el SAPI de Microsoft tenía ya muy desarrollado un modelo de lenguaje para dicho idioma, por lo cual se decidió tomar esa herramienta y adaptar mas bien su modelo de lenguaje a las necesidades específicas de la aplicación, en la cual se limitó el vocabulario a comandos compuestos por dos o más palabras en un orden específico relacionados con el control que un usuario de la silla de ruedas necesita, al cerrar el modelo de lenguaje para que reconozca solamente los comandos definidos se logra aumentar la confiabilidad del sistema. La aplicación demostró ser independiente del hablante y no requerir de entrenamientos previos, puesto que cada persona para realizar las pruebas solo tuvo que empezar a pronunciar los comandos definidos e inmediatamente el sistema los empezó a reconocer exitosamente.

- Se validó la respuesta del sistema de reconocimiento de comandos de voz en español, visualizando los resultados en matrices de confusión sobre las que se calcularon parámetros de eficiencia correspondientes a la exactitud global, a la sensibilidad, la especificidad, la precisión y la medida F1 de los diferentes comandos pronunciados en las pruebas. Cada uno de los parámetros de eficiencia mencionados se obtuvieron de manera diferenciada en la respuesta que el sistema dio para interlocutores de género femenino e interlocutores de

género masculino en tres rangos de niveles de ruido que abarcan desde un entorno en silencio hasta un entorno con alta percepción de ruido donde establecer una comunicación con otra persona se dificulta. Según los resultados encontrados, no hay diferencias significativas en la respuesta del sistema según género del interlocutor (Tabla 13). Por su parte, al realizar el análisis en los tres rangos de nivel de ruido se encontró que a medida que el ruido aumenta, la respuesta del sistema de reconocimiento va disminuyendo pero en muy poca proporción, comportándose casi igual con diferencias de no mas de 1,25% en sus parámetros de eficiencia para las pruebas en los dos primeros rangos de ruido (de 35 dB(A) hasta 55 dB(A) y de 60 dB(A) hasta 72 dB(A)) y mostrando un poco mas de cambio entre la prueba en un entorno en silencio y la realizada con el ruido más alto (73 dB(A) hasta 85 dB(A)), con diferencias no mayores a 5% (presentado para la sensibilidad) en sus parámetros de eficiencia (Tabla 19). La mínima diferencia en los resultados entre ambientes, muestra que el sistema tiene un muy buen comportamiento aún en lugares con presencia de ruido de 85 dB(A), siendo su respuesta 100% óptima en el ambiente de prueba más silencioso.

- Se desarrolló una aplicación computacional para el reconocimiento independiente del hablante de comandos de voz en español, adaptando un modelo general de lenguaje existente en dicho idioma a una gramática específica para el usuario de una silla de ruedas que desea comunicarse y transmitir órdenes de control desde la misma, limitando así al reconocedor para escuchar sólo los comandos de interés con lo que se mejora la respuesta de la aplicación. Las pruebas realizadas muestran que el sistema es robusto ya que su exactitud casi no se disminuye bajo condiciones de ruido en el entorno obteniéndose en el caso de mayor ruido (prueba con 73 dB(A) hasta 85 dB(A)) una exactitud de 98,82% frente a la exactitud de 100% resultante en un entorno en silencio (Tabla 19), así mismo los demás parámetros de eficiencia presentan disminuciones muy leves con el incremento del ruido en el ambiente. Los errores del sistema de reconocimiento sobre los comandos se presentaron en mayor medida por omisión que por sustitución entre los mismos. Las pruebas también indicaron que el sistema responde de manera muy similar con interlocutores de ambos géneros (Tabla 13). Las pruebas complementarias realizadas para comprobar los efectos de alzar la voz al momento de pronunciar los comandos bajo condiciones de ruido en el ambiente (efecto Lombard) demostraron que como restricción importante para que el sistema de reconocimiento funcione asertivamente, se debe siempre pronunciar con un nivel de voz similar al que sin esfuerzo un

interlocutor emite en un ambiente en silencio, independientemente del ruido presente al momento de pronunciar los comandos (sección 4.2.5).

6. RECOMENDACIONES

- Si un sistema de reconocimiento de voz solo va a detectar ciertos comandos de interés, la mejor opción será construir una gramática que limite al reconocedor para que solo distinga esas palabras. De esta manera se incrementará la precisión del reconocedor, el gasto de memoria de reduce, se evita una respuesta equívoca entre palabras fonéticamente similares, entre otras ventajas.
- Con el sistema de reconocimiento de comandos de voz en español que se construyó para el presente proyecto, se puede comenzar a integrar la respuesta al reconocimiento con los diferentes módulos de Hardware que se están realizando en el grupo de Automática de la UAM para la toma de signos vitales, órdenes de domótica y movimiento de la silla inicialmente.
- Se debe evitar subir la voz al momento de pronunciar los comandos en entornos con nivel de ruido apreciable (efecto Lombard), ya que esto interfiere enormemente en la respuesta del reconocedor, los cambios al subir la voz afectan no solo a la sonoridad, sino también a factores como el tono, el rango y la duración del sonido de las sílabas, dando como resultado una disminución importante en todos los parámetros de eficiencia del reconocedor al aumentar los casos de errores por omisión o por sustitución entre los comandos.
- Como muestra los resultados obtenidos en la prueba 4.3.1, se recomienda para formar los comandos, utilizar palabras que tengan más de una sílaba y que no presenten problemas comunes en la pronunciación como es el caso de la doble r, ya que éste tipo de palabras obtuvieron una mayor probabilidad de reconocimiento erróneo, así como los casos con valores de confidencia más bajos.

7. BIBLIOGRAFÍA

[1] D. Jurafsky and J.H. Martin, *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*, 2nd ed.: Pearson Prentice Hall, 2009.

[2] X. Huang and L. Deng, "An Overview of Modern Speech Recognition," in *Handbook of Natural Language Processing*, 2nd ed.: Chapman & Hall/CRC, 2010, ch. 15 (ISBN: 1420085921), pp. 339-366.

[3] Organización Mundial de la Salud and Banco Mundial. (2011) Informe mundial sobre la discapacidad. [Online]. http://www.who.int/disabilities/world_report/2011/es/index.html

[4] DANE - Dirección de Censos y Demografía. (2010, Mar) Registro de Localización y Caracterización de las Personas con discapacidad. [Online]. http://www.dane.gov.co/index.php/poblacion-y-demografia/discapacidad

[5] L.I. Mejía and M. Narváez, *Trascendiendo la limitación física: Enfoque y manejo integral de algunos problemas potencialmente invalidantes*. Manizales, Caldas: Editorial Andina, 1996.

[6] L. Quintero, *Trauma: Abordaje inicial en los servicios de urgencias*, 3rd ed. Cali: Publicaciones Salamandra, 2005.

[7] (2014, Oct) Jazzy Electric Wheelchairs. [página web]. [Online]. http://www.jazzy-electric-wheelchairs.com/

[8] (2014, Oct) Merits. [página web]. [Online]. http://www.merits.com.tw/index.php

[9] (2014, Oct) Quickie. [página web]. [Online]. http://www.quickiepower.com/index.asp?locale=en_GB

[10] C.S.L. Tsui et al., "EMG-based hands-free wheelchair control with EOG attention shift detection," in *IEEE Int'l Conf. Robotics and Biomimetics. (ROBIO 2007)*, 15 - 18 Dec. 2007, pp. 1266-1271.

[11] S. Yathunanthan et al., "Controlling a Wheelchair by Use of EOG Signal," in *4th Int'l Conf. Information and Automation for Sustainability. (ICIAFS 2008)*, 12-14 Dec. 2008, pp. 283-288.

[12] I. Iturrate, J. Antelis, and J. Minguez, "Synchronous EEG brain-actuated wheelchair with automated navigation," in *IEEE Int'l Conf. Robotics and Automation. (ICRA '09)*, 12-17 May. 2009, pp. 2318-2325.

[13] Z. Hu et al., "A novel intelligent wheelchair control approach based on head gesture recognition," in *2010 Int'l Conf. Computer Application and System Modeling. (ICCASM)*, 22-24 Oct. 2010, pp. V6-159-V6-163.

[14] M.E. Lund et al., "Inductive tongue control of powered wheelchairs," in *Annual International Conference of the IEEE. Engineering in Medicine and Biology Society. (EMBC)*, Aug. 31 2010-Sept. 4 2010, pp. 3361-3364.

[15] A. Murai et al., "Voice activated wheelchair with collision avoidance using sensor information," in *ICCAS-SICE, 2009*, 18-21 Aug. 2009, pp. 4232-4237.

[16] M. Fezari and A.-E. Khati, "New speech processor and ultrasonic sensors based embedded system to improve the control of a motorised wheelchair," in *Design and Test Workshop. IDT 2008. 3rd International*, 20-22 Dec. 2008, pp. 345-349.

[17] M.T. Qadri and S.A. Ahmed, "Voice Controlled Wheelchair Using DSK TMS320C6711," in *Int'l Conf. Signal Acquisition and Processing. (ICSAP 2009)*, 3-5 April. 2009, pp. 217-220.

[18] M. Fezari, M. Bousbia-Salah, and M. Bedda, "Voice and Sensor for More Security on an Electric Wheelchair," in *2nd International Conference on Information and Communication Technologies. (ICTTA '06)*, 2006, pp. 854-858.

[19] Z. Yuhong, "Controlling the Intelligent Wheelchair by distinguishing emotion, illness and Environment," in *2nd Int'l Conf. Artificial Intelligence, Management Science and Electronic Commerce. (AIMSEC)*, 8-10 Aug. 2011, pp. 2016-2019.

[20] A. Skraba et al., "Prototype of speech controlled cloud based wheelchair platform for disabled persons," in *Embedded Computing (MECO), 2014 3rd Mediterranean Conference*, 15-19 June. 2014, pp. 162,165.

[21] M. Fezari, S. Lemboub, and M.S. Boumaza, "VR-Stamp with DSP-TMS320C6711 for hand-free voice-driven monitoring robots navigation," in *Information Technology and e-Services (ICITeS), 2012 International Conference*, 24-26 March. 2012, pp. 1,7.

[22] Inc Sensory. (2000) Voice Direct™ 364 Data Book. [Online]. http://www.datasheetarchive.com/Voice-

Direct-364-datasheet.html

[23] C. Aruna, A. Dhivya Parameswari, M. Malini, and G. Gopu, "Voice recognition and touch screen control based wheel chair for paraplegic persons," in *Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference*, 6-8 March 2014, pp. 1,5.

[24] (2014, Oct) Images Scientific Instruments Inc. [Online]. http://www.imagesco.com/speech/SR-07.pdf

[25] A. Murai et al., "Elevator available voice activated wheelchair," in *The 18th IEEE International Symposium. Robot and Human Interactive Communication.(RO-MAN 2009)*, Sept. 27 2009-Oct. 2 2009, pp. 730-735.

[26] H. Kokubo et al., "Embedded Julius: Continuous Speech Recognition Software for Microprocessor," in *IEEE 8th Workshop. Multimedia Signal Processing*, 3-6 Oct. 2006, pp. 378-381.

[27] (2014, Oct) Open-Source Large Vocabulary CSR Engine Julius. [Online]. http://julius.sourceforge.jp/en_index.php?q=index-en.html

[28] A. LEE. (2010, May) The Julius book. Edition 1.0.3 - rev.4.1.5. [Online]. http://julius.sourceforge.jp/en_index.php

[29] M.A.M. Abushariah et al., "Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools," in *2010 Int'l Conf. Computer and Communication Engineering. (ICCCE)*, 11-12 May. 2010, pp. 1-6.

[30] Y. Wang and X. Zhang, "Realization of Mandarin continuous digits speech recognition system using Sphinx," in *2010 International Symposium. Computer Communication Control and Automation. (3CA)*, 5-7 May. 2010, pp. 378-380.

[31] (2014, Oct) CMU Sphinx - Open Source Toolkit For Speech Recognition. [Online]. http://cmusphinx.sourceforge.net/

[32] T. Imbiriba et al., "GMM and kernel-based speaker recognition with the ISIP toolkit," in *2004 14th IEEE Signal Processing Society Workshop. Machine Learning for Signal Processing*, Sept. 29 2004-Oct. 1 2004, pp. 371-380.

[33] The Institute for Signal and Information Processing. (2014, Oct) ISIP toolkit. [Online]. http://www.isip.piconepress.com/projects/speech/software/

[34] H.Q. Nguyen, V.L. Trinh, and T.D. Le, "Automatic Speech Recognition for Vietnamese Using HTK System," in *2010 IEEE RIVF International Conference. Computing and Communication Technologies, Research, Innovation, and Vision for the Future. (RIVF)*, 1-4 Nov. 2010, pp. 1-4.

[35] B.A.Q. Al-Qatab and R.N. Ainon, "Arabic speech recognition using Hidden Markov Model Toolkit(HTK)," in *2010 International Symposium. Information Technology. (ITSim)*, 15-17 June. 2010, pp. 557-562.

[36] (2014, Oct) HTK Speech Recognition Toolkit. [Online]. http://htk.eng.cam.ac.uk/

[37] (2014, Oct) Microsoft Speech API. [Online]. http://msdn.microsoft.com/en-us/library/ee125077(v=VS.85).aspx

[38] (2014, Oct) VoxForge. [Online]. http://www.voxforge.org

[39] (2015, Sept) Apple Inc. [Online]. https://support.apple.com/es-es/HT4992

[40] (2015, Oct) Google. [Online]. http://www.google.com/search/about/features/01

[41] M. Nishimori, T. Saitoh, and R. Konishi, "Voice controlled intelligent wheelchair," in *SICE, 2007 Annual Conference*, 17-20 Sept. 2007, pp. 336-340.

[42] J.C. Martinez and J.L. Ramirez, "Diseño y construcción de un módulo automático controlado por voz adaptable a una silla de ruedas convencional," in *Segundo Congreso Internacional de Ingeniería Mecatrónica ISSN: 1234-1234*, Colombia, 2009.

[43] O.I. Higuera, "Diseño e implementación de un prototipo de reconocimiento de voz basado en modelos ocultos de markov para comandar el movimiento de una silla de ruedas en un ambiente controlado," in *XII Simposio de Tratamiento de Señales, Imágenes y Visión artificial*, Colombia, 2007.

[44] W. Acosta, M. Sarria, and L. Duque, "Implementación de una metodología para la detección de comandos de voz utilizando HMM," *Revista de Investigaciones - Universidad del Quindío*, vol. 23(1), pp. 64-70, 2012.

[45] (2014, Oct) discapacidadonline.com. [Online]. http://www.discapacidadonline.com/ingenieros-colombianos-universidad-los-llanos-crean-silla-ruedas-controlada-voz.html

[46] D. Ruiz, *C#: La Guía Total del Programador - Manuales Users.code*.: MP Ediciones, 2005.

[47] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing, a guide to theory, algorithm and system development*, (ISBN: 0130226165), Ed.: Prentice Hall, 2001.

[48] L. Lamel and J. Gauvain, "Speech Recognition," in *The Oxford Handbook of Computational Linguistics*.: Oxford University Press, 2003, ch. 16, pp. 305-322.

[49] T. L. Floyd, *Fundamentos de Sistemas Digitales*, 9th ed. Madrid: Pearson Educación S.A, 2006.

[50] F.J. Hernando, "Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidosos," Tesis Doctoral, Universitat Politécnica de Catalunya. Departamento de Teoria de la señal y comunicaciones, ISBN: 9788469136911, 1993.

[51] J.A. Morales, "Técnicas de reconocimiento robusto de la voz basadas en el pitch," Tesis Doctoral, Universidad de Granada. Departamento de Teoría de la Señal Telemática y Comunicaciones, ISBN: 9788469493441, 2011.

[52] J. Tebelskis, "Speech Recognition using Neural Networks," doctoral dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, May 1995.

[53] L.A. Ripoll, "Verificación de hablante basado en Dynamic Time Warping," *Ingeniería & Desarrollo. Universidad del Norte*, no. 3_4. (ISSN electrónico: 2145-9371), pp. 111-127, 1998.

[54] J.D.V. Peña, "Contribuciones al reconocimiento robusto de habla," Tesis Doctoral, Universidad Carlos III de Madrid. Departamento de Teoría de la Señal y Comunicaciones, 2007.

[55] J.A. Gámez and J.M. Puerta, *Sistemas Expertos Probabilísticos*. España: Ediciones de la Universidad de Castilla - La Mancha, 1998.

[56] R. Brown. MSDN Magazine, "Exploring New Speech Recognition And Synthesis APIs In Windows Vista". [Online]. http://msdn.microsoft.com/en-us/magazine/cc163663.aspx

[57] Microsoft. (2014, Oct) System.Speech Programming Guide for.NET Framework. [Online]. http://msdn.microsoft.com/en-us/library/hh361633(v=office.14).aspx

[58] (2015, Feb) Speech Recognition Grammar Specification. [Online]. http://www.w3.org/TR/speech-grammar

[59] G. E Dahl et al., "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30-42, Enero 2012.

[60] Microsoft. (2014, Nov) Centro de noticias. [Online]. http://www.microsoft.com/es-es/news/Press/2014/Nov14/Microsoft-codigo-abierto.aspx

[61] Microsoft. (2014, Oct) Introducing Computer Speech Technology. [Online]. http://msdn.microsoft.com/en-us/library/ms870025

[62] AECOR - Asociación Española para la Calidad Acústica, "Guía y procedimiento de medida del ruido de actividades en el interior de edificios.," 2011. [Online]. http://www.caib.es/sacmicrofront/archivopub.do?ctrl=MCRST147ZI116097&id=116097

[63] Organización Mundial de la Salud. (1999) Guías para el ruido urbano. [Online]. http://www.bvsde.paho.org/bvsci/e/fulltext/ruido/ruido2.pdf

[64] (2015, Marzo) myNoise™.net. [Online]. http://mynoise.net/NoiseMachines/cafeRestaurantNoiseGenerator.php?c=0&l=49949494559494947459

[65] S.Planet, "Reconocimiento afectivo automático mediante el análisis de parámetros acústicos y lingüisticos del habla espontánea," Tesis Doctoral, Escola Tècnica superior d'Enginyeria Electrònica i informàtica La Salle. Departamento Comunicacions i Teoria del Senyal, Barcelona,.

[66] N. Kawarazaki and T. Yoshidome, "Remote control system of home electrical appliances using speech recognition," in *Automation Science and Engineering (CASE), 2012 IEEE International Conference*, 20-24 Aug. 2012, pp. 761,764.

8. ANEXOS

Anexo 1. Instrumento de recolección de pruebas

Edad:
Género:
Ambiente en el que se realiza la prueba (rango en dB(A)):

Nota: Los espacios con X indican que si se obtuvo un reconocimiento exitoso, en caso de no reconocer ningún comando se deja el espacio en blanco. En caso de reconocer erroneamente otro comando se escribe el comando erróneo.

Comandos relacionados con el movimiento de la silla						
Comandos	**Intentos (válidos o no válidos)**				**Comando reconocido erróneamente**	**Total Aciertos**
Mover adelante						0
Mover atrás						0
Mover izquierda						0
Mover derecha						0
Mover parar						0
Mover lento						0
Mover rápido						0

Comandos relacionados con órdenes de domótica						
Comandos	**Intentos (válidos o no válidos)**				**Comando reconocido erróneamente**	**Total Aciertos**
Prender luz pasillo						0
Prender luz cocina						0
Apagar luz pasillo						0
Apagar luz alcoba						0
Abrir puerta entrada						0
Cerrar cortina alcoba						0
Abrir cortina entrada						0
Cerrar puerta alcoba						0

Comandos relacionados con la toma de signos vitales						
Comandos	**Intentos (válidos o no válidos)**				**Comando reconocido erróneamente**	**Total Aciertos**
Capturar electro						0
Capturar temperatura						0
Medir electro						0
Medir oxígeno						0
Tomar presión						0
Tomar pulso						0

Comandos relacionados con el desplazamiento por la pestañas						
Comandos	**Intentos (válidos o no válidos)**				**Comando reconocido erróneamente**	**Total Aciertos**
ver programas						0
ver silla						0
Ver inicio						0
Ver domótica						0
Ver salud						0
ver configuración						0

Comandos relacionados con el control de los botones						
Comandos	**Intentos (válidos o no válidos)**				**Comando reconocido erróneamente**	**Total Aciertos**
Voz detener						0
voz activar						0

Comandos relacionados con destinatarios de correo electrónico						
Comandos	**Intentos (válidos o no válidos)**				**Comando reconocido erróneamente**	**Total Aciertos**
Correo médico						0
Correo vecino						0
Correo hermano						0

Comandos relacionados con la apertura de aplicaciones						
Comandos	**Intentos (válidos o no válidos)**				**Comando reconocido erróneamente**	**Total Aciertos**
Aplicación calculadora						0
Aplicación fecha						0
Aplicación notas						0

Tabla A 1: Instrumento de recolección de pruebas.

Anexo 2. Matriz de confusión de la prueba en un ambiente ruidoso para comprobar el efecto Lombard en la disminución de la calidad de respuesta del sistema de reconocimiento.

Prueba con efecto Lombard
Rango 73 dB(A) hasta 85 dB(A)

Entrada \ Salida	Mover adelante	Mover atrás	Mover izquierda	Mover derecha	Mover parar	Mover lento	Mover rápido	Prender luz pasillo	Prender luz cocina	Apagar luz pasillo	Apagar luz alcoba	Abrir puerta entrada	Cerrar cortina alcoba	Abrir cortina entrada	Cerrar puerta alcoba	apagar luz entrada	Capturar electro	Capturar temperatura	Medir electro	Tomar Electro	Medir oxígeno	Tomar presión	Tomar pulso	Tomar oxígeno	ver programas	ver silla	Ver inicio	Ver domótica	Ver salud	ver configuración	Ver pruebas	Voz detener	voz activar	Correo médico	Correo vecino	Correo hermano	Aplicación calculadora	Aplicación fecha	Aplicación notas	No reconocido	Total Comandos
Mover adelante	10																			1																				1	12
Mover atrás		6																		1																				5	12
Mover izquierda			10																																					2	12
Mover derecha				4																2																				6	12
Mover parar					11																																			1	12
Mover lento						2														9																				1	12
Mover rápido							9																	1																2	12
Prender luz pasillo								0		3										1				1																7	12
Prender luz cocina									2																															10	12
Apagar luz pasillo										11																														1	12
Apagar luz alcoba											10					2																									12
Abrir puerta entrada												6		2																										4	12
Cerrar cortina alcoba													6																											6	12
Abrir cortina entrada														12																											12
Cerrar puerta alcoba															0																									12	12
apagar luz entrada																12																									12
Capturar electro																	9			3																					12
Capturar temperatura																		10																						2	12
Medir electro																			7	3																				2	12
Tomar Electro																			1	11																					12
Medir oxígeno																					10			2																	12
Tomar presión																				3		4																		5	12
Tomar pulso																						1	10	1																	12
Tomar oxígeno																								12																	12
ver programas																									0																12	12
ver silla																				2						2															8	12
Ver inicio																											8													4	12
Ver domótica																												0												12	12
Ver salud																													9											3	12
ver configuración																						1								8											3	12
Ver pruebas																				4					2						0										6	12
Voz detener																				2											1										9	12
voz activar																																11									1	12
Correo médico																				8													2								2	12
Correo vecino																				4				5										0							3	12
Correo hermano																				4															5						3	12
Aplicación calculadora																																				11				1	12
Aplicación fecha																																					10			2	12
Aplicación notas																																						10		2	12

Tabla A 2: Matriz de confusión de la prueba en un ambiente ruidoso para comprobar el efecto Lombard en la disminución de la calidad de respuesta del sistema de reconocimiento.